


Systematic Review and Meta-Analysis of Stay-Play-Talk Interventions for Improving Social Behaviors of Young Children

Journal of Positive Behavior Interventions
2023, Vol. 25(1) 65–77
© Hammill Institute on Disabilities 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1098300720983521
jpbi.sagepub.com


Jennifer R. Ledford, PhD¹ , and James E. Pustejovsky, PhD²

Abstract

Stay-play-talk (SPT) is a peer-mediated intervention that involves training peer implementers to stay in proximity to, play with, and talk to a focal child who has disabilities or lower social competence. This systematic review and meta-analysis investigated the contexts in which SPT interventions have been conducted, the methodological adequacy of the research assessing its effects, and the outcomes for both peer implementers and focal children. Studies have primarily occurred in inclusive preschool settings during free play activities, with researchers serving as facilitators. Average effects were positive and substantial for both peer implementers and focal children, although considerable heterogeneity across studies was observed. Additional research is needed to determine what peer implementer and focal child characteristics moderate intervention success, what modifications are needed for children who have complex communication needs, and optimal procedural variations (e.g., group size, training time).

Keywords

peer-mediated procedures, competence, social relationships/interactions

Peer-related social competence is an important area of growth for all young children, but especially for children who have or are at risk for social delays (Odom et al., 2008). Social competence in preschool is associated with a wide range of positive outcomes over time (e.g., academic achievement and mental health; Jones et al., 2015), and children who are enrolled in classrooms with more positive peer interactions in preschool have lower rates of problem behavior in elementary school (Spivak & Farran, 2016). Given the inverse relation between social competence and problem behavior, as well as the negative outcomes associated with problem behavior (i.e., suspension and expulsion; U.S. Department of Education Office for Civil Rights, 2014), it is clear that social competence is critical in early childhood contexts.

Adult support of peer relations is associated with children's socially competent behavior (Irvin et al., 2015). Early childhood practitioners can choose from a range of evidence-based practices designed to improve skill acquisition, social-emotional development, and peer relationships for young children. Two broad categories of interventions for improving social competence include teacher-directed interventions and peer-mediated interventions (PMIs). Several teacher-directed strategies, such as the use of systematic prompting and reinforcement, have been found to be effective for improving social interactions for young

children. However, peer-mediated strategies may lead to considerable changes in social behaviors of children's peers, perhaps resulting in more durable, long-term changes in children's interactions and relationships.

Peer-mediated interventions are characterized as treatments delivered by children who are close in age to the recipients, usually classmates or siblings. Peer-mediated strategies involve identifying one or more peers (sometimes referred to as *confederates* or *implementing peers*) with higher social competence and training them to interact with another child or children (sometimes referred to as *target* or *focal children*) who have a disability and/or lower social competence (Odom & Strain, 1984). Often, interactions occur in the presence of adult supports (e.g., prompting) for the confederate but without direct, adult-facilitated treatment for focal participants (Goldstein et al., 2007; Odom & Strain, 1984). Instead, treatment for the focal participant is delivered by the peer implementer. The Division for Early

¹Vanderbilt University, Nashville, TN, USA

²The University of Wisconsin-Madison, USA

Corresponding Author:

Jennifer R. Ledford, Vanderbilt University, Peabody Box 228, Nashville, TN 37203, USA.

Email: Jennifer.ledford@vanderbilt.edu

Action Editor: Daniel Maggin

Childhood of the Council for Exceptional Children identified the use of PMI as a recommended practice for improving engagement and learning for young children with or at risk for disabilities (Division for Early Childhood, 2014).

Odom and Strain (1984) classified PMIs to improve social behaviors as proximity interventions (teaching peers to stay close by), prompt and reinforce interventions (teaching peers to prompt and reinforce focal children's engagement in specific behaviors), and peer-initiation interventions (teaching peers to make overtures to focal children). Early iterations of peer-initiation interventions focused on complex, responsive interactions between typically developing children and children with disabilities, and they resulted in positive changes in child behavior (Goldstein et al., 2007). Although these peer-initiation interventions were effective, they required considerable adult prompting for the peer implementer, which limited feasibility (i.e., teachers reported that they were difficult to implement alongside typical teaching responsibilities). In response, Goldstein and colleagues (2007) developed stay-play-talk (SPT) interventions. SPT was established to simplify demands on peers by eliminating complex requirements and identifying just three behaviors designed to improve social interactions between "buddies"—staying near a focal child, playing with that child, and talking about their play. SPT was designed to reduce the need for frequent teacher prompts, and SPT behaviors were identified as likely to result in positive reciprocal peer interactions because they did not always involve requisite responding on the part of the focal participant (Goldstein et al., 2007). When SPT interventions are conducted, children are taught to stay, play, and talk with specific buddies during *training sessions* and then are asked to use those behaviors during a non-training activity (*measurement sessions*). Although SPT has been identified as a PMI, the extent to which confederates versus confederates *and* focal children have been trained is unclear. For example, if only confederates are trained, the intervention is PM, and changes in focal child behavior can be attributed to peer implementer behavior. However, if both confederates and focal children are trained, the intervention includes both PM and adult-mediated components and changes in child behavior may be due to peer behavior, adult-implemented supports, or both.

Several recent reviews have evaluated the extent of evidence for PMIs (Chan et al., 2009; Chapin et al., 2018; Watkins et al., 2015). These reviews identified that PMIs often included modeling, prompting, and reinforcement and that they were generally effective for improving communication and social behaviors for children with autism. Despite mostly positive outcomes, authors noted that it was difficult to identify *for whom and under what conditions* PMIs were optimally effective (e.g., prerequisite skills, training procedures). The aforementioned PMI reviews included only children with autism spectrum disorders (ASD). While children with ASD have weaknesses in peer-related social

competence, many children with social competence support needs do not have ASD. Thus, it may be reasonable to review specific PMIs for children with disabilities more generally.

In addition to unanswered questions regarding participant characteristics, the previous reviews also had some methodological shortcomings. First, two restricted their assessment to published sources. This potentially increases the risk of bias in estimating intervention effects due to selective publication practices (Shadish et al., 2016; Tincani & Travers, 2018). That is, estimates from published sources only may result in inaccurate, systematically biased estimates. In addition, none provided estimates of magnitude using meta-analytic procedures and appropriate, contemporary effect sizes.

Finally, no separate syntheses of the impacts of SPT interventions for peer implementers or focal children have been published and none of the aforementioned reviews assessed outcomes separately for SPT as a unique PMI category. Thus, a synthesis of effect sizes for SPT as a single, well-defined variation of PMI is needed.

Research Questions

The purpose of this review was to describe the use of SPT for improving social behaviors of young children, to characterize study quality and rigor, and to estimate the average and distribution of effects of SPT interventions on the behaviors of both implementing peers and focal children. Research questions guiding the review were as follows: In studies in which SPT condition was compared with a non-SPT condition:

- RQ1.** What are the characteristics of participants, settings, implementers, and dependent and independent variables?
- RQ2.** To what extent are studies characterized by adequate rigor and quality?
- RQ3.** What are average effect sizes of using SPT, and what are the distributions of effects for focal children and implementing peers?

Method

Inclusion Criteria and Search Procedures

Inclusion criteria assessed for potentially eligible studies are shown in Table S1 (online supplemental materials) and include use of an SPT intervention, implementation with young children in free operant contexts, and assessment with an experimental design. The search was conducted in December 2018 using the following search string: "stay play talk" OR "stay, play, talk" OR "stay-play-talk" OR "peer budd*" OR "buddy skill*" in the PsycINFO and ProQuest Dissertations and Theses databases. In June 2020, an additional search was conducted using the ERIC database; no additional relevant sources were identified. To

decrease the likelihood of publication bias impacting our findings, we included both published and unpublished sources that met minimum rigor requirements. Interobserver agreement was assessed for 100% of steps (e.g., database search, forward search) and potential sources (i.e., identified articles) for all search procedures using point-by-point agreement. In all instances, agreement on inclusion was calculated by dividing the number of agreements by total number of potential articles (i.e., total number generated by the search). The primary assessor was an expert with a doctorate in special education who conducted all search procedures. Two secondary assessors were master's students in education programs, both of whom were completing coursework and fieldwork to become certified behavior analysts; each independently conducted some of the search procedures conducted by the primary assessor for the purpose of assessing reliability. These students were trained in person by the first author to identify relevant independent variables for inclusion in the review by reviewing inclusion and exclusion criteria using behavioral skills training (written and verbal descriptions, modeling with a non-included study, and practice with non-included studies followed by written feedback). After completion of these steps, coders began coding included studies; there were no exit criteria for training.

We conducted screening in several steps to reduce complexity for secondary coders. We first identified all studies that met all but the design inclusion criterion in Table S1 in the supplemental materials (i.e., we screened in studies even if they used non-eligible designs, such as A–B designs). From this search, 16 sources were identified, two of which were unpublished duplicates of published studies ($n = 14$ unique sources; 99.7% agreement). Backward (examining and assessing for inclusion sources cited in the 14 identified articles) and forward searches (identifying via Google Scholar and assessing for inclusion sources that later cited the 14 identified articles) were conducted; no additional sources were identified (100% agreement). To identify sources that may not have been indexed in the original searches due to recent publication, we conducted two searches of sources published in 2018 or 2019. First, we hand searched five journals relevant to early childhood, autism, or speech-language pathology (*Focus on Autism and Other Developmental Disabilities*, *Journal of Autism and Developmental Disorders*, *Journal of Early Intervention*, *Journal of Speech, Language, and Hearing Research*, and *Topics in Early Childhood Special Education*) for those years, including articles published online but not yet assigned to an issue. These journals were selected based on the first author's expertise because they (a) focused on populations likely to be included in SPT studies, (b) commonly published intervention research related to these populations, and (c) frequently published single-case design studies. Two additional articles were identified (99.6% agreement). Finally, a search was conducted using Google

Scholar for sources published “since 2018.” Each term in our search string was entered separately (with all three “stay play talk” terms resulting in identical results); one additional unpublished dissertation was identified (99.2% agreement; see “Screening Results”; <https://osf.io/3wv47/>). In a second step, we screened SPT studies to determine whether they met design criteria; this was done for two reasons. First, it reduced task complexity. Second, it allowed us to easily conduct forward and backward searches on a larger body of SPT-relevant literature (i.e., even studies with low rigor may be likely to include citations to related work). Seven sources were excluded because they used nonexperimental single-case designs and one was excluded because it compared SPT with and without certain treatment components (see “Excluded Sources”; <https://osf.io/3wv47/>). Thus, a total of nine sources were included.

Descriptive Coding

Descriptive data were coded in the following domains to describe for whom and under what conditions SPT interventions were implemented: general information, setting and adult facilitators, focal participants, peers, dependent variables, training, measurement sessions, and social validity. All data were coded by the primary assessor described above; 100% of data were double coded by one of eight secondary assessors. The secondary assessors were all graduate students in a university special education department and were either board certified behavior analysts (BCBAs) or were in the process of completing coursework and experiences required to sit for the BCBA exam. Prior to coding, all secondary coders were required to read the coding manual, code two excluded studies, calculate agreement with the primary coder, and participate in discrepancy discussions if any disagreements occurred. Coders also participated in discrepancy discussions throughout the coding process; if a disagreement occurred, it was calculated as a disagreement, and coders discussed whether it was due to error (in which case the accurate code was used in the final coding file) or disagreement (in which case coders discussed the discrepancy and determined via consensus which was correct; this consensus code was used in the final coding file). Average agreement was 96.7%; see <https://osf.io/3wv47/> for more detail.

General information coded included publication year, status (peer-reviewed or not), number of eligible focal participants, design type, and number of opportunities to detect a potential functional relation (e.g., in a study with three participants in a multiple baseline across participants design, one opportunity to detect a potential functional relation was present for each measured dependent variable). Setting and adult facilitator information coded included type of setting (e.g., inclusive, home, clinic), training context, measurement context (e.g., free play, playground), and

adult facilitators for training and measurement sessions (e.g., teacher, parent, researcher). Participant information coded included total number, educational placement, age, gender, race/ethnicity, and disability information. We also coded whether inclusion criteria for participation were reported; we recorded descriptive information about pre-intervention language, social-emotional, and play skills when reported by authors. Each dependent variable in a given source was coded; information included total number of dependent variables per source, who emitted the behavior (e.g., focal, peer, other), to whom behaviors were directed (e.g., identified buddies, any child in classroom), behavior category (e.g., broad social interactions, SPT strategy use), and recording system (e.g., interval, count). For intervention focus, we coded types of training provided, training participants (e.g., peers only, buddy groups, whole-class), methods used (e.g., role-play, feedback), and dosage (e.g., size of group, number and duration of training sessions). For social validity, we coded whether data were reported regarding preference of participants or stakeholders, normative comparisons (e.g., whether social interactions reached typical levels exhibited by a nonparticipant), or blind raters.

Rigor Coding

We evaluated rigor using the Single Case Analysis and Review Framework (SCARF; Ledford et al., 2016), which allowed for an assessment of outcomes in relation to rigor, with results of both plotted on a scatterplot. SCARF requires yes/no coding for 44 codes in seven quality/rigor domains (reliability, fidelity, data sufficiency, social and ecological validity, participant descriptions, condition descriptions, generalization measurement, maintenance measurement) and three outcome domains (primary, generalization, maintenance). SCARF is one of a number of assessment tools for single-case studies. It provides a more thorough assessment of quality and rigor when compared with the What Works Clearinghouse (WWC; 2019) guidelines, which were designed to serve as a minimum standard rather than to provide descriptive information about quality and rigor (Zimmerman et al., 2018). Moreover, SCARF gives greater weight to internal validity considerations (rigor, reliability, fidelity, data sufficiency) than to other variables that may not be critical for determining whether a functional relation exists. The SCARF is also more complete than the WWC standards (e.g., includes considerations related to fidelity and social validity) and more specific (e.g., contains operationalized definitions and ratings) than the Council for Exceptional Children Standards for Evidence-Based Practices in Special Education (Zimmerman et al., 2018).

SCARF results in quality and rigor scores of between 0 and 4, with scores above 2 generally considered to be adequate.

Because SPT results in measurement of generalized responding (i.e., primary data come from different contexts than training for generalized behaviors; Ledford & Gast, 2018), we did not code generalization measurement or outcomes. Generally, outcome scores of 3 or 4 are consistent with the presence of a functional relation. Scoring was conducted by the primary coder and one secondary coder as described earlier, with overall agreement of 94.3% (agreement by code available via “SPT IOA”; <https://osf.io/3wv47/>).

Visual Analysis of Outcomes

Selecting outcomes. Two independent coders, a doctoral level BCBA who is an expert in single-case design and a graduate student who was a BCBA with considerable experience conducting single-case studies, conducted visual analysis for all eligible outcomes in each design in all sources (e.g., if two dependent variables were measured in the context of a multiple baseline design, we evaluated the presence of a functional relation separately for each variable). Some studies included multiple potential functional relations for each participant (e.g., authors measured multiple behaviors for a single participant), and some included separate potential functional relations for two peers who simultaneously implemented the intervention for the same focal participant. Visual analysis was conducted for 35 potential functional relations in nine sources (1–14 potential functional relations in each source).

Analysis. Visual analysis was conducted by evaluating level, trend, and variability within and between conditions; immediacy of change and degree of overlap between conditions; and the consistency of change across potential demonstrations of effect (Ledford & Gast, 2018; WWC, 2019). Functional relations were affirmed when consistent changes in level occurred between conditions (i.e., change occurred when and only when conditions changed), with an immediate or near-immediate change in level with or without continuing therapeutic trends during intervention, with at least three demonstrations of effect in a single design. We evaluated the presence of functional relations on a 0 to 4 scale, consistent with typical SCARF procedures. A score of 0 indicated no behavior change (or counter therapeutic change, which did not occur in this review); scores of 1 to 2 indicate inconsistent behavior change with one or more non-effects, and scores of 3 to 4 indicate consistent behavior change with no non-effects. Generally, scores of 3 and 4 are consistent with determinations that a functional relation exists with a 3 indicating that some effects are “weak” (e.g., delayed, with some overlapping data, or smaller in magnitude than effects in other tiers) and 4 indicating consistent and unequivocal positive effects.

Effect Sizes and Meta-Analysis

Selecting outcomes. For purposes of effect size calculation, we selected one dependent variable for each participant in each source, purposefully chosen to maximize comparability across studies. For peer participants, most researchers measured a behavior we categorized as a “broad social interaction”; thus, we used this variable to estimate effect sizes for peers to maximize the comparability of outcomes across studies. These interactions were defined as social behaviors that were general (e.g., interactions, communicative acts) rather than specific (e.g., narrowly defined behaviors such as staying in proximity or responding to greetings). For focal participants, most studies reported only one outcome measure. For studies that measured more than one variable for focal participants, we used the one that was categorized as a broad social interaction. Other measured variables included use of SPT strategies, joint attention, and more narrowly defined social interactions; none of these variables were measured consistently across multiple studies. Because these variables were usually components of broad social interaction, and because very few studies that measured any one of them, we judged it would be reasonable to exclude social behaviors not defined as broad social interactions from the meta-analysis while including the broader interaction measure.

Log response ratios (LRR). The LRR effect size is the natural log of proportionate change in the mean level between conditions (Pustejovsky, 2018). This metric is appropriate for direct measures of behavior, as commonly used in single-case designs (Ledford & Gast, 2018). Furthermore, the LRR has been shown to be less susceptible to change based on procedural variations (e.g., session length and measurement system) and has fewer requirements (i.e., it does not require multiple participants within a given source) compared with other effect sizes such as non-overlap measures or the between-case standardized mean difference (Pustejovsky, 2018, 2019b). To facilitate interpretation, the LRR can be converted into percentage change, a familiar and accessible outcome metric. Two versions of LRR are available: one for syntheses where most studies measure behaviors expected to increase during intervention conditions (LRRi) and one for syntheses where most studies measure behaviors expected to decrease (LRRd). We used the LRRi because all dependent variables were social behaviors expected to improve.

For the purposes of calculating effect sizes, a graduate student extracted all available data from all eligible graphs from included studies using PlotDigitizer (2015). Obtained values were corrected under two conditions. First, if the value was not possible given the range of the scale (e.g., -0.1 or 100.1 on a percentage scale), we corrected the obtained value to be equal to the nearest possible value

(e.g., 0 or 100). Second, if there were 21 or fewer possible values, we rounded obtained values to the nearest possible value. For example, if there were 20 intervals per session, resulting in possible values of 0, 5, 10, and so on, a value of 4.9 would be corrected to 5; if the value was a count and the axis ranged from 0 to 10, extracted values were rounded to the nearest whole number (given that a part of a behavior cannot occur during a session). Values for dependent variables that had more than 21 possible values (e.g., 50 intervals per session, rates per minute) were not corrected given that the more possible values that existed, the more difficult it is to ensure accuracy in corrections.

We used the SingleCaseES package (Pustejovsky & Swan, 2018) in the R statistical environment (R Core Team, 2018) to calculate LRR estimates for each focal participant; in most cases, the estimate was from a single A–B comparison in a multiple baseline or multiple probe design across participants. In some cases, we calculated an average effect across two A–B comparisons in a withdrawal design or across multiple social partners in a multiple baseline design across peer implementers. We calculated corresponding effects for implementing peers (i.e., one peer effect for each focal participant effect). When multiple peers implemented the intervention simultaneously (i.e., during the same sessions) for the same focal participant and their data were presented separately (i.e., two data paths), we calculated a single average effect size for each set of peers paired with a single focal participant. Average effects were estimated by calculating separate LRR values for each A–B comparison and then taking the arithmetic mean; standard errors for the average effects were calculated as described in Pustejovsky (2018).

Meta-analysis. We then synthesized the LRR effect sizes using two types of multilevel random effects meta-analysis models. First, we estimated separate meta-analysis models for focal child and for peer implementer behaviors. Second, we estimated a multivariate meta-analysis model for the joint distribution of focal child and peer implementer effects. All multilevel models included random effects at the level of the study and at the level of the participant. The standard deviation of the study-level random effects describes the degree of heterogeneity across studies, as might arise due to variation in implementation, outcomes, settings, or participant inclusion criteria; the standard deviation of the participant-level random effects describes the extent of heterogeneity across participants within a study. The multivariate models included additional parameters capturing the correlation between focal child and peer implementer errors, at both the study level and the participant level. For estimating the overall average effects across studies and participants, we used cluster-robust variance estimation methods to account for potential inaccuracy of the effect size standard errors due to auto-correlation

(Pustejovsky, 2018). Before conducting meta-analyses, we examined the distribution of effect size estimates for outliers. One study (Milam, 2018) consisted of three outlying estimates, represented substantively different participants than all other studies (children at risk for rather than with disabilities), only measured a dependent variable for peers that did not represent a broad social interaction, and had near-zero baseline levels (a context in which LRR does not perform well); thus, this study was excluded from the meta-analysis.

To facilitate interpretation of the meta-analysis results, we report results both on the LRR metric and after translating into the metric of percentage change. In the LRR metric, we report the estimated average effect from each model, the corresponding standard error, and a 95% confidence interval for the average effect. As measures of heterogeneity, we also report the estimated study-level and participant-level standard deviations from the multilevel random effects model. At each level, larger standard deviations indicate greater variability (less consistency) in outcomes. In the percentage change metric, we report the estimate average effect and corresponding 95% confidence interval. To translate the study-level and participant-level heterogeneity estimates, we calculate 67% prediction intervals (PIs; Borenstein et al., 2017) for each level of the model. The study-level PI describes the range of *average* effect size that one would expect to observe in 67% of further studies, drawn from the same population as the included studies. The participant-level PI describes the range of individual effect sizes that one would expect to observe in 67% of participants in further studies. At each level, wider intervals indicate greater variability in anticipated outcomes.

We used the metafor (Viechtbauer, 2010) and clubSandwich (Pustejovsky, 2019a) packages in the R statistical environment (R Core Team, 2018) for estimating meta-analysis models and obtaining cluster-robust variances, respectively. Raw data and the R script for replicating the meta-analysis are available via the Supplemental Materials (“stay-play-talk_meta-analysis”; <https://osf.io/3wv47/>).

Results

Below, we provide data to answer each of the following research questions: **(RQ1)**: What are the characteristics of participants, settings, adult facilitators, dependent variables, and independent variables? **(RQ2)**: To what extent are studies characterized by adequate rigor and quality? **(RQ3)**: What are average effect sizes of using SPT, and what are the distributions of effects?

Descriptive Analysis

General information. Nine included sources primarily used time-lagged designs across participants ($n = 6$) or withdrawal designs ($n = 2$), with one source using a multiple baseline design across social partners. One source included one ineligible design and one eligible design (one concurrent and one nonconcurrent design; Goldstein et al., 1997), and two sources included ineligible and eligible participants (Milam, 2018, some potential participants did not receive intervention; Tsao & Odom, 2006, one participant did not meet age criteria)—results are reported only for eligible participants and designs. Most studies (seven of nine) included measurement of both focal participant behaviors and implementing peer behaviors; the remaining two sources included measurement of only focal participant behavior. The total number of potential functional relations was 35 (see Table 1).

Participants. There were 25 focal and 31 individual peer participants (including siblings). One study (Laushey & Heflin, 2000) included all non-focal participants in a classroom as peers. When reported individually, the average age was 56 months for focal participants (range = 36–90 months; reported in eight of nine sources) and 59 months for peers (range = 36–101 months; reported in five of nine sources). When gender was reported, most focal participants ($n = 18$; 72%) and implementing peers ($n = 19$; 63%) were reported to be male. Race was unreported for 54% of focal participants and 60% of peer implementers. When reported individually, seven focal participants and seven peer implementers were identified as White, three focal participants and five peer implementers were identified as Black, and one each was identified as Asian. In one study, six White children, one Native American child, and one child of Middle Eastern ethnicity participated, but we could not ascertain which children were focal participants versus implementing peers. No participants were reported as being Hispanic. The most common disability reported for focal participants was ASD ($n = 11$); other common disabilities were developmental delay ($n = 6$), intellectual disability ($n = 5$), and hearing/vision impairments ($n = 2$). Three focal participants were identified as being *at risk* for social-emotional delays (Milam, 2018). Two peers were identified as also having disabilities.

Inclusion criteria. Inclusion criteria for focal participants were unreported in five of nine sources. The only inclusion criteria reported in more than one study were enrollment or attendance ($n = 2$) and low social skills ($n = 3$). Others were minimum language, social, and play skills; age; ASD diagnosis; and consent ($n =$ one each). Of the eight sources that included individual peer participants, inclusion criteria for implementing peers varied widely and were unreported

Table 1. Study and Measurement Characteristics and Visual Analysis Outcomes for Focal Participants and Peer Implementers, by Design.

First author	Design	Behavior type	Measurement	Visual analysis (SCARF score)	No. of effect sizes
Focal participants					
Baldwin	MB → SP	Broad SI	Count	FR (4)	1
Barber	MB → P	Initiations	Count	No FR (0) ^a	—
	MB → P	Responses	Count	No FR (0) ^a	—
English	MB → P	Broad SI	Count	No FR (1) ^a	3
	MP → P	Broad SI	Count	FR (4)	4
Goldstein	MP → P	Broad SI	Count	No FR (1)	4
Kim	MB → P	Broad SI	Interval ^b	FR (3)	3
Laushey	A–B–A–B	Specific SI	percentage	FR (4)	1
	A–B–A–B	Specific SI	percentage	FR (4)	1
Milam	MP → P	Play	Duration	FR (4)	—
	MP → P	Broad SI	Count	No FR (1)	3
Severini	A–B–A–B	Broad SI	Count	No FR (0)	1
	A–B–A–B	Broad SI	Count	No FR (0)	1
Tsao	MB → P	Broad SI	Interval ^b	No FR (1)	4
	MB → P	Specific SI	Interval ^b	No FR (0)	—
Peer implementers					
Barber ^a	MB → P	Initiations	Count	No FR (1) ^a	—
	MB → P	Responses	Count	No FR (1) ^a	—
	MB → P	Broad SI	Count	FR (3) ^a	3
English	MP → P	Broad SI	Count	FR (4)	4
Goldstein	MP → P	Broad SI	Count	FR (4)	4
Kim	MB → P	Broad SI	Interval ^b	FR (4)	3
Milam	MP → P	Strategy use	Duration	FR (4)	3
Severini	A–B–A–B	Broad SI	Count	No FR (0)	1
	A–B–A–B	Stay	MTS	FR (3)	—
	A–B–A–B	Play	MTS	FR (4)	—
	A–B–A–B	Talk	MTS	FR (0)	—
	A–B–A–B	Broad SI	Count	FR/No FR (3/0)	1 ^c
	A–B–A–B	Stay	MTS	FR/FR (3/4)	—
	A–B–A–B	Play	MTS	FR/FR (4/4)	—
	A–B–A–B	Talk	MTS	No FR/No FR (0/0)	—
Tsao	MB → P	Broad SI	Interval ^b	No FR (0)	3

Note. SCARF = Single Case Analysis and Review Framework; No. of effect sizes = number of effect sizes calculated from this design (see the online supplemental materials for effect sizes); MB = multiple baseline; SP = social partner; SI = social interactions; FR = functional relation; P = participant; MP = multiple probe; MTS = momentary time sampling.

^aLow-rigor study. ^bNot enough information provided to determine specific interval system used. ^cFor four designs in Severini, data from two peers were combined into a single effect size; two separate functional relation determinations were made.

in two sources. In reporting sources, the most common criteria were presence of adequate social skills ($n = 6$) and demographic or convenience variables rather than skill-related behaviors ($n = 4$). Other criteria were adequate play skills ($n = 2$), language skills ($n = 2$), compliance ($n = 2$), cognitive skills ($n = 2$), empathy/maturity skills ($n = 2$), and likely to play in same area or interact with focal participant ($n = 2$).

Setting and adult facilitators. Interventions were typically implemented in inclusive or general education classrooms ($n = 6$), with two conducted in the home with siblings, and

one conducted in a clinic. Of the six sources reporting classroom settings, four were in preschool and two were in elementary school settings (Baldwin, 2013; Laushey & Heflin, 2000). Most sources ($n = 6$) reported that measurement occurred in free play or centers; two reported that measurement occurred during free play plus other activities (playground, snack, large group), and one reported that measurement occurred during other activities (lunch, physical education [PE], art). In one study, the classroom teacher and researchers both facilitated training, whereas in seven studies researchers facilitated training (facilitator type unreported in one study). In one study each, teachers and

paraprofessionals alone or with researchers facilitated measurement session procedures (e.g., rewards); in the other studies, researchers alone facilitated procedures.

Dependent variables. Sources included measurement of one to six different dependent variables, for a total of 25 different variables across studies, in 35 different experimental designs. About half of the measured dependent variables ($n = 13$) were emitted by focal participants, whereas others ($n = 12$) were emitted by peer participants or by both participants and nonparticipants (e.g., children in classrooms who were not assigned “buddies,” $n = 2$). Seven of nine studies reported at least one outcome for both peer and focal participants. Table 1 shows the variable type and measurement system for each variable, by experimental design. Broad social interaction was the most common variable type ($n = 16$). Count was the most common measurement system ($n = 17$), with duration or an estimate of duration (momentary time sampling) used in 11 designs.

Training. All studies included two components: (a) *training*, during which children were taught to engage in SPT behaviors, and (b) *measurement sessions*, during which outcomes were measured. Authors in six sources reported training only implementing peers, with no training for focal participants. Other authors reported whole-group training provided to all children in the classroom ($n = 1$), whole-group training plus buddy-group training (triads of implementing peers plus focal participants; $n = 1$), or buddy-group training plus separate training for implementing peers regarding use of augmentative or alternative communication (AAC) devices.

During training sessions, a variety of teaching methods were used by adult facilitators. In all sources, an explanation of or rationale for use of stay-play-talk behaviors was provided. Most ($n = 7$) explicitly reported using role-play or practice opportunities, whereas fewer explicitly reported using modeling ($n = 5$) or feedback ($n = 3$). The only additional component reported was the use of electronic or paper-based stories ($n = 3$). The mean and median total number of training sessions was 5 (range = 1–13); when reported, the duration per session ranged from 10 to 20 min. When both number and duration of sessions was reported, the mean training time was 73 min (range = 40–100 min; when a range was reported by a source, we used the midpoint to calculate this average; when only a maximum was reported, we excluded from calculation).

Measurement sessions. There was an average of nine measurement sessions per intervention condition (average range across sources = 5–16), for average total duration of 107 min. Presence of prompting and praise were not well-described across sources; authors in seven sources explicitly reported prompting peer implementers, while presence

of prompting was unclear in two sources. Authors for only one source explicitly reported prompting of focal participants and one explicitly did not prompt; remaining seven sources were not explicit. Similarly, the majority of studies were not explicit about whether peer implementers or focal children received rewards (e.g., praise, tokens) during the sessions. Authors often reported that peer implementers received some reward following the session ($n = 6$), whereas only two reported that focal children received a similar reward. However, authors in only one study explicitly reported that rewards were not provided (for focal participants only); thus, in three studies for peer implementers and six for focal participants, the nature of post-session reinforcement was unclear. For studies conducted in homes or clinics ($n = 3$), buddies were constant across sessions. For studies in classrooms, buddies were constant across measurement occasions in three sources and alternating or variable in three sources.

Social validity. Most ($n = 7$) sources included at least one measure of social validity. Five reported data regarding the acceptability of the intervention by non-implementing stakeholders. One reported data regarding acceptability by child participants. Five sources used secondary raters to assess child performance via video; in four sources, raters were explicitly identified as being naive to condition. All social validity data were positive; that is, SPT was reported as acceptable and resulting in favorable outcomes.

Rigor Analysis

The average quality/rigor score across studies was 2.4 for peer implementers and 2.3 for focal participants (possible range of 0–4; actual range of 1.0–3.2); see Figure 1 in the online supplemental materials. Only one source contributed scores lower than 2.0 (Barber et al., 2016); in this study, researchers reported collecting interobserver agreement data only during intervention conditions and did not collect fidelity data. Averages across domains were generally above 2.0, with the exception of maintenance measurement, with less than one third of designs including measurement of maintained behavior change, and social validity (primarily resulting from low scores for all designs from Severini et al., 2019, a study in which social validity was not assessed).

Outcomes

Visual analysis. Visual analysis was conducted for 35 designs (from nine sources). For focal participants, data in six of 15 designs (40%) demonstrated a functional relation (scores of 3 or 4). When a functional relation was not identified, it was primarily due to inconsistent effects (e.g., one nonresponder in a multiple baseline across participants design) or failure

Table 2. Results of Univariate Multilevel Meta-Analyses.

Outcome	LRR				Percentage change			
	Average effect (SE)	95% confidence interval	Study-level SD	Participant-level SD	Average effect	95% confidence interval	67% prediction interval (study)	67% prediction interval (participant)
Including all studies with children with disabilities as focal participants								
Peers	1.30 (0.16)	[0.89, 1.71]	0.31	0.27	267	[144, 453]	[139, 464]	[90, 530]
Focal participants	1.12 (0.27)	[0.48, 1.77]	0.74	0.06	208	[61, 487]	[43, 561]	[45, 502]
Excluding Severini et al. (2019)								
Peers	1.34 (0.18)	[0.84, 1.85]	0.37	0.00	282	[131, 533]	[156, 471]	[156, 471]
Focal participants	1.29 (0.23)	[0.74, 1.84]	0.57	0.04	263	[109, 530]	[100, 558]	[100, 559]

Note. Severini et al. (2019) was the only study including focal participants who used augmentative communication devices. All meta-analysis excluded Milam (2018). LRR = log response ratios.

to control for threats to internal validity (e.g., increasing baseline data that continued to increase during intervention conditions). For peer implementers, data in 12 of 20 designs (65%) demonstrated a functional relation, 11 of which were associated with designs with high quality/rigor scores. Six designs had high quality/rigor scores but did not demonstrate positive outcomes (Figure 1 in the Supplemental Materials).

Effect size estimates. LRR effect size estimates are reported in Table S2 in the online supplemental materials. Peer estimates ranged from 0.45 to 6.99, and focal child estimates ranged from -2.63 to 2.59. There were three very large outlying effect sizes for peer implementers, all from Milam (2018). In these cases, the peers displayed near-zero levels of behavior during baseline—a context where the LRR is not an appropriate measure of effect (Pustejovsky, 2018). In addition, this study represented the only assessment of SPT with focal participants who were identified as being at risk for social-emotional delays rather than having identified disabilities. Given substantive and methodological differences, we excluded the results of this study from the meta-analyses (although we retained the study in all other analyses). Remaining peer estimates ranged from 0.45 to 2.37. The two lowest estimates of focal child behavior change (see density and scatter plots; “stay-play-talk_meta-analysis”; <https://osf.io/3wv47/>) came from the same study (Severini et al., 2019); these participants were the only two participants in the group of studies who were reported to use AAC devices. Because we hypothesized, based on these estimates, that outcomes for children using AAC might differ in magnitude, we conducted meta-analyses both with and without including these effect sizes.

Comparisons between visual analysis and effect sizes. For peer participants, effect sizes in designs associated with functional relations were on average larger than those associated with designs that did not demonstrate a functional relation (2.03

vs. 1.13). The exception is one comparison for Severini et al. (2019) with no functional relation but an effect size of 2.01; this discrepancy is due to near-zero baseline levels (as mentioned previously, a situation in which LRR tends to perform less well). For focal participants, effect sizes in designs associated with functional relations were larger than those associated with no functional relations (1.37 vs. 0.77). Designs in which functional relations were demonstrated but with relatively smaller magnitudes had *consistent* magnitudes and designs with no functional relations that included A–B comparisons with relatively larger magnitudes had *highly variable* magnitude values across comparisons in a single design (e.g., across tiers).

Meta-analysis. Because results from the multivariate and univariate analyses were similar, we report the results from the simpler univariate model in Table 2 (with and without excluding Severini et al., 2019). Results of the multivariate analyses are available in the online supplemental materials (“stay-play-talk_meta-analysis,” <https://osf.io/3wv47/>). The average effect for peer behaviors was $LRR_i = 1.30$ (95% CI = [0.89, 1.71]), which corresponds to an increase in social behaviors of 267% (95% CI = [144%, 453%]). As shown in Table 2, estimates are similar for the model excluding Severini et al. (2019; average $LRR_i = 1.34$, 282% increase). The 67% PI estimates for the latter model are 156% and 471% at both study and participant levels, respectively, indicating substantial variability in expected average effects from new studies.

The average effect for focal child behaviors was $LRR_i = 1.12$ (95% CI = [0.48, 1.77]), which corresponds to an increase in social behaviors of 208% (95% CI = [61%, 487%]). As shown in Table 2, estimates are slightly higher and more precise for the model excluding Severini et al. (2019; effect size = 1.29, 263% increase). Notably, there was substantial between-study variation in average effects for focal child behavior, as indicated by the wide study-level PI (67% PI = [100%, 558%]), but little within-study

variation in effects (estimated participant-level $SD = 0.04$ from the model excluding Severini et al.).

Discussion

The purpose of this review was to evaluate the existing evidence regarding the use of and outcomes for SPT interventions, which were first assessed more than 20 years ago. SPT intervention studies included primarily preschool-age participants, with focal participants often having autism, developmental delays, or intellectual disabilities. Sessions were often conducted in free play, by researchers, and often included measurement of a generalized social behavior such as interactions. Training frequency and duration varied widely and procedures for measurement sessions (e.g., reinforcement) were often not well-described.

Results of both visual analysis and meta-analysis indicate positive outcomes for implementing peers and focal participants, for improving broad social interactions during free play activities in preschool classrooms and homes. Two studies, both occurring in elementary school classrooms, reported only focal participant results; because we do not have data for peer implementers, information about the impact of SPT in elementary school settings is incomplete. Comparisons between functional relation determinations and effect sizes indicated that effect size estimates were generally in line with functional relation determinations, such that larger magnitude estimates were associated with the presence of a functional relation. Visual analysis allowed for identification of functional relations and detection of data patterns that might be potentially problematic during effect size estimation, whereas the meta-analysis allowed for separate estimates of magnitude and variability in outcomes for focal participants and implementing peers.

In terms of visual analysis, a larger proportion of functional relations were identified for implementing peers than for focal participants. Similarly, in terms of effect sizes, the magnitude of behavior change was larger, on average, for implementing peers. This is not surprising, given that (a) focal participants often received less training than peers, (b) the intervention period was relatively brief (e.g., about 3 hr, including both training and measurement sessions), and (c) focal participant effects were dependent on peer implementation. There were no instances in which there was no functional relation for peers but a demonstrated relation for focal participants. This provides evidence that changes in peer behaviors are an active ingredient for the intervention received by focal participants. This was generally measured as changes in broad social behaviors emitted by peers rather than procedural fidelity, *per se*. Thus, although changes in peer behavior appear to be critical, it is less clear to what extent changes in each targeted behavior (staying, playing, and talking) are associated with positive outcomes for focal participants.

No study including multiple dependent variables for individual participants demonstrated functional relations for each variable. This pattern of findings underscores the importance of identifying which analyses are confirmatory and exploratory for single-case studies, and which variables were used to make experimental decisions. For example, it may be that peer implementer behavior was most often used to make decision about condition changes, resulting in data patterns that were less interpretable for focal participants (e.g., baseline data with an increasing trend prior to intervention initiation). This suggests that effects may differ based on experimental procedures rather than response to intervention—but the basis of procedural decisions was rarely described in the study sources.

Results of the meta-analysis also indicated variation in the effectiveness of SPT—particularly variation across studies—for both peer and focal participant behavior. The estimated average effects were clearly positive, corresponding to substantial improvements of 267% and 208% from baseline levels of peer and focal participant behavior. However, the degree of positive behavior change was less consistent. Some variability in implementation (e.g., training peers only vs. buddy groups together) and measurement (e.g., counting the number of occurrences vs. measuring or estimating duration) was apparent. It is possible that these procedural variations or unreported differences contribute to heterogeneity in the magnitude of behavior change. Other potential sources of heterogeneity were obscured due to inadequate reporting—for example, the degree to which prompting and reinforcement was provided during sessions was often unreported and inclusion criteria were often missing or unrelated to behaviors likely to be needed to successfully implement or benefit from the intervention. Given additional research and more thorough reporting practices, it may be possible to explore the relations via meta-regression in the future.

In terms of training, procedures were nearly always aligned with PM procedures—that is, peer implementers were most often trained without focal participants. Although the addition of direct social skills instruction for children who are less socially competent seems likely to result in positive outcomes, two studies that included direct instruction in buddy groups (i.e., with focal participants) failed to demonstrate a functional relation for focal participant behaviors for three of four opportunities to do so (Milam, 2018; Severini et al., 2019). Because these two studies were excluded from the meta-analysis altogether (Milam, 2018) or in one model (Severini et al., 2019), the reported outcomes primarily represent SPT implemented as a PMI rather than a combined or whole-class intervention. The extent to which focal participants received prompting or reinforcement during measurement sessions is unclear due to lack of explicit reporting of assistance provided during these sessions. Additional research and better reporting of *in situ* assistance provided to peers and focal participants is needed.

After excluding studies without three potential demonstrations of effect, nearly all sources met rigor/quality standards; this is discrepant from other bodies of work, even those including only peer-reviewed studies (Zimmerman et al., 2018; Zimmerman & Ledford, 2017). Rigor scores were unrelated to publication status—the lowest-scoring source was from a peer-reviewed article and the highest-scoring source was an unpublished dissertation. These findings provide further evidence that excluding single-case design studies based on publication status may not accurately screen against lower-quality research (Ledford & Gast, 2018).

Limitations

Because our search procedures required authors to use all three terms associated with the intervention (stay, play, talk), it is possible that some relevant interventions may not have been included. However, because SPT is a specific “named” intervention, it seems unlikely that an eligible study would have both failed to use the name and failed to cite or be cited by any included studies. Although we attempted to use the most similar dependent variables across studies, no standardized outcome measures are associated with SPT, and researchers used a variety of methods for measuring social interactions (e.g., count, interval systems). Thus, synthesizing outcomes across studies comes with the usual “comparing apples to oranges” problem associated with meta-analyses. As a result, some of the observed heterogeneity in effect sizes may be due to variation in operationalization and measurement of dependent variables. In addition, because we selected the journals based on author expertise, we may have introduced bias; additional studies may be available but unidentified. Finally, poor reporting by primary study authors prevented careful analysis of procedural variations that may have occurred between studies (e.g., prompting, reinforcement).

Future Directions

Although results of the current review suggest SPT may be an effective intervention for improving social behaviors of children with low social competence and their peers, additional research is needed to answer questions critical to understanding variability in outcomes. First, information is needed regarding inclusion criteria for implementing peers—that is, we need evidence about what skills make a child likely to be an effective implementer. More than 30 years ago, Odom and Strain (1984) suggested that clear and explicit criteria for choosing peer models were needed. Data from this review suggest that minimal progress has been made on this front, at least in relation to SPT interventions. Relatedly, few studies reported fidelity of peer implementation, a common problem in PMIs (Chan et al., 2009).

Additional work is needed to establish what level of fidelity is needed to promote behavior change for focal children and to help practitioners select peers who are likely to implement SPT with sufficiently high fidelity.

Inclusion criteria for focal children were minimally useful for determining which children might benefit from SPT. For example, the intervention appeared somewhat less successful for the two participants for whom use of AAC devices was explicitly reported, despite the fact that other studies did not explicitly describe a vocal communication requirement for inclusion. It is difficult to interpret this finding because these two participants also received intervention as part of the same study (Severini et al., 2019) and thus some procedural or contextual factor unique to this study could be the underlying cause of the outlying values. Regardless, additional research with children with complex communication needs (CCN) is needed. Due to limited available evidence, it is not clear what modifications in SPT procedures might result in optimal outcomes for children with CCN. For example, a common treatment for children who use AAC devices is aided AAC modeling, wherein adults model appropriate communication on a child’s device (Biggs et al., 2018). Because children may “model” verbal language without explicit teaching (i.e., talk), but may not model AAC use, SPT interventions for children who use AAC devices may need to incorporate additional peer training and reinforcement related to incorporation of aided language.

Other participant characteristics might also moderate treatment effects. For example, authors in two recent studies (Osborne et al., 2019; Severini et al., 2019) suggest that responsivity to peer overtures and functional play repertoires might be relevant prerequisite skills for focal children; teacher-directed interventions used in addition to PMIs such as SPT may be required for these children. More research is needed to determine the extent to which these types of characteristics influence response to intervention. Authors of future studies should report rationales and procedures for participant selection to assist researchers and practitioners to identify for whom and under what conditions SPT is likely to lead to behavior change.

In addition to improved participant information, research is needed to determine effects of (a) adult prompting versus other reminders during measurement sessions (Osborne et al., 2019), (b) single versus multiple peer buddy assignments, (c) rotating versus constant peer buddy assignments, and (d) training group size and composition. Research is also needed to establish the effects of SPT when indigenous facilitators provide training and the extent to which they can and will implement procedures. Data are also needed to confirm whether peer implementers maintain behavior change over time, given lack of maintenance data and relatively short intervention durations reported. Finally, data are needed to determine whether outcomes for focal children maintain or

improve, in the presence or absence of peer implementation maintenance.

Conclusion

Even given very short training sessions and practice opportunities, SPT was demonstrated to result in consistently positive changes in social interactions for peer implementers and smaller, less consistent changes for social interactions for focal children. Although SPT should be considered a promising practice for improving social interaction behaviors in free play contexts, considerable variability in outcomes across studies suggests that some contextual or procedural factors may influence the magnitude of behavior change.

Acknowledgments

The authors thank Kate T. Chazin for her thoughtful feedback on an earlier version of this paper.

Declaration of Conflicting Interests

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

Funding

This work 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.

ORCID iD

Jennifer R. Ledford  <https://orcid.org/0000-0002-2392-7103>

Supplemental Material

Supplemental material for this article is available on the *Journal of Positive Behavior Interventions* website with the online version of this article.

References

References marked with an asterisk indicate studies included in the systematic review

- *Baldwin, B. (2013). *Peer-mediated intervention and technology: Mobile device application use in implementation of peer buddy training* [Unpublished dissertation]. Florida State University, Tallahassee.
- *Barber, A. B., Saffo, R. W., Gilpin, A. T., Craft, L. D., & Goldstein, H. (2016). Peers as clinicians: Examining the impact of Stay Play Talk on social communication in young preschoolers with autism. *Journal of Communication Disorders, 59*, 1–15.
- Biggs, E. E., Carter, E. W., & Gilson, C. B. (2018). Systematic review of interventions involving aided AAC modeling for children with complex communication needs. *American Journal on Intellectual and Developmental Disabilities, 123*, 443–473.
- Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*, 5–18.
- Chan, J. M., Lang, R., Rispoli, M., O'Reilly, M., Sigafoos, J., & Cole, H. (2009). Use of peer-mediated interventions in the treatment of autism spectrum disorders: A review. *Research in Autism Spectrum Disorders, 3*, 876–889.
- Chapin, S., McNaughton, D., Boyle, S., & Babb, S. (2018). Effects of peer support interventions on the communication of preschoolers with autism spectrum disorder: A systematic review. *Seminars in Speech and Language, 39*, 443–457.
- Division for Early Childhood. (2014). *DEC recommended practices*. Retrieved April 4, 2019, from <https://divisionearlychildhood.egnyte.com/dl/tgv6GUXhVo>
- *English, K., Goldstein, H., Shafer, K., & Kaczmarek, L. (1997). Promoting interactions among preschoolers with and without disabilities: Effects of a buddy skills-training program. *Exceptional Children, 63*, 229–243.
- *Goldstein, H., English, K., Shafer, K., & Kaczmarek, L. (1997). Interaction among preschoolers with and without disabilities: Effects of across-the-day intervention. *Journal of Speech, Language, and Hearing Research, 40*, 33–48.
- Goldstein, H., Schneider, N., & Thiemann, K. (2007). Peer-mediated social communication intervention: When clinical expertise informs treatment development and evaluation. *Topics in Language Disorders, 27*, 182–199.
- Irvin, D. W., Boyd, B. A., & Odom, S. L. (2015). Adult talk in the inclusive classroom and the socially competent behavior of preschoolers with autism spectrum disorder. *Focus on Autism and Other Developmental Disabilities, 30*, 131–142.
- Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health, 11*, 2283–2290.
- *Kim, T. (2010). *Sibling-implemented intervention for improving social interaction skills of young children who have difficulties with socialization* [Unpublished dissertation]. University of Kansas, Lawrence.
- *Laushey, K. M., & Heflin, L. J. (2000). Enhancing social skills of kindergarten children with autism through the training of multiple peers. *Journal of Autism and Developmental Disorders, 30*, 183–193.
- Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology*. Routledge.
- Ledford, J. R., Lane, J. D., Zimmerman, K. N., Chazin, K. T., & Ayres, K. A. (2016, April). Single case analysis and review framework (SCARF). <http://ebip.vkcsites.org/scarf/>
- *Milam, M. E. (2018). *Stay-play-talk with preschoolers: Programming for generalization* [Unpublished dissertation]. Vanderbilt University, Nashville, TN.
- Odom, S. L., McConnell, S. R., & Brown, W. H. (2008). Social competence of young children: Conceptualization, assessment, and influences. In W. H. Brown, S. L. Odom, & S. R. McConnell (Eds.), *Social competence of young children: Risk, disability, and intervention* (pp. 3–29). Paul H. Brookes.

- Odom, S. L., & Strain, P. S. (1984). Peer-mediated approaches to promoting children's social interaction: A review. *American Journal of Orthopsychiatry*, *54*, 544–557.
- Osborne, K., Ledford, J. R., Martin, J., & Thorne, K. (2019). Component analysis of stay, play, talk interventions with and without self-monitored group contingencies. *Topics in Early Childhood Special Education*, *39*, 5–18.
- Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112.
- Pustejovsky, J. E. (2019a). *clubSandwich: Cluster-Robust (Sandwich) variance estimators with small-sample corrections* [R package version 0.3.4]. <https://cran.r-project.org/web/packages/clubSandwich/index.html>
- Pustejovsky, J. E. (2019b). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*, 217–235.
- Pustejovsky, J. E., & Swan, D. M. (2018). *SingleCaseES: A calculator for single-case effect size indices* [R package version 0.4.0.9999]. <https://github.com/jepusto/SingleCaseES>
- R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- *Severini, K. E., Ledford, J. R., Barton, E. E., & Osborne, K. (2019). Implementing stay-play-talk with children who use AAC. *Topics in Early Childhood Special Education*, *38*, 220–233.
- 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*, 656–673.
- Spivak, A., & Farran, D. (2016). Predicting first graders' social competence from their preschool classroom interpersonal context. *Early Education and Development*, *6*, 735–750.
- Tincani, M., & Travers, J. (2018). Publishing single-case research design studies that do not demonstrate experimental control. *Remedial and Special Education*, *39*, 118–128.
- *Tsao, L., & Odom, S. L. (2006). Sibling-mediated social interaction intervention for young children with autism. *Topics in Early Childhood Special Education*, *26*, 106–123.
- U.S. Department of Education Office for Civil Rights. (2014). *Civil rights data collection: Data snapshot: School discipline*. <http://ocrdata.ed.gov/Downloads/CRDC-School-Discipline-Snapshot.pdf>
- Wiechthbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.
- Watkins, L., O'Reilly, M., Kuhn, M., Gevarter, C., Lancioni, G. E., . . . Lang, R. (2015). A review of peer-mediated social interventions for students with autism in inclusive settings. *Journal of Autism and Developmental Disorders*, *45*, 1070–1083.
- What Works Clearinghouse. (2019) *Procedures and standards handbook. Version 4.1*. <https://ies.ed.gov/ncee/wwc/Protocols#procedures>
- Zimmerman, K. N., & Ledford, J. R. (2017). Beyond ASD: Evidence for the effectiveness of social narratives. *Journal of Early Intervention*, *39*, 199–217.
- Zimmerman, K. N., Pustejovsky, J. E., Ledford, J. R., Barton, E. E., Severini, K. E., & Lloyd, B. P. (2018). Single-case synthesis tools II: Comparing quantitative outcome measures. *Research in Developmental Disabilities*, *79*, 65–76.