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

Jennifer R. Ledford
Peabody Box 228
Vanderbilt University
Nashville, TN 37203
Jennifer.ledford@vanderbilt.edu

James E. Pustejovsky
1912 Speedway
M/S D5800
University of Texas at Austin
Austin, TX 78712
pusto@austin.utexas.edu

Acknowledgement: We would like to thank Kate T. Chazin for her thoughtful feedback on an earlier version of this paper.

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.

Abstract

Stay-play-talk (SPT) is a peer-mediated intervention which 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 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).

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

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, McConnell, & Brown, 2008). Social competence in preschool is associated with a wide range of positive outcomes over time (e.g., academic achievement and mental health; Jones, Greenberg, & Crowley, 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, 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, Boyd, and Odom, 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 (TD) and peer-mediated (PM) interventions. Several TD strategies, such as the use of systematic prompting and reinforcement, have been found to be effective for improving social interactions for young children. However, PM strategies may lead to considerable changes in social behaviors of children's *peers* (Odom & Strain, 1986), perhaps resulting in more durable, long-term changes in children's interactions and relationships.

PM interventions have been characterized as treatments delivered by children who are close in age to the recipients. PM strategies generally 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. 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, Schneider, & Thiemann, 2007; Odom & Strain, 1984). Instead, treatment for the focal participant is delivered by the peer. The Division for Early Childhood of the Council for Exceptional Children (DEC) identifies the use of PM intervention as a recommended practice for improving engagement and learning for young children with or at risk for disabilities (DEC, 2014).

Odom and Strain (1984) classified PM interventions 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 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. In response, Goldstein and colleagues 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 PM intervention, the extent to which confederates versus confederates *and* focal children have been trained is unclear; that is, the intervention may be PM (implemented by confederates) or a combination intervention (involving confederate implementation plus TD strategies for focal children).

Several recent reviews have evaluated the extent of evidence for PM interventions in general. Some have restricted assessments to participants with autism spectrum disorders (ASD; Chan et al., 2009; Chapin, McNaughton, Boyle, & Babb, 2018; Watkins et al., 2015). Although children with ASD have weaknesses in peer-related social competence, many children with social competence support needs do not have ASD. In addition, previous reviews have sometimes restricted inclusion to peer-reviewed studies (Chan et al., 2009; Watkins et al., 2015); such restrictions could increase the risk of bias in assessments of intervention effects due to selective publication practices (Shadish, Zelinsky, Vevea, & Kratochwill, 2016; Tincani & Travers, 2018). Finally, no syntheses of the impacts of SPT interventions for peer implementers or focal children have been published. The previous reviews instead included a wide range of intervention strategies for teaching peers to engage in many different behaviors; it may be helpful to have specific effect size estimates for single, well-defined variations, such as SPT.

Research Questions

The purpose of this review is to describe the use of SPT for improving social behaviors of young children 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: In studies in which an SPT condition is compared to a non-SPT condition, (1) What are the characteristics of participants, settings, adult facilitators, dependent variables, and independent variables? (2) To what extent are studies characterized by adequate rigor and quality? And (3) What are average treatment effects associated with SPT, and what are the distribution of effects for focal children and implementing peers?

Methods

Inclusion Criteria

To be included, sources had to (a) include an eligible study assessing the effects of an SPT intervention; (b) include children aged birth to age 8 years, 11 months who had identified disabilities, delays, or risks for disabilities or delays; (c) include measurement in free operant contexts, during which children had opportunities to choose to engage in “stay” behaviors (e.g., if measurement occurred during a table activity in which children were not allowed to leave, it was excluded); (d) include direct measurement of at least one social behavior for peers or focal participants; and (e) include assessment of effects of an SPT intervention condition in relation to a non-SPT condition, using an experimental design. To qualify as an SPT intervention, explicit information had to be provided that skills taught included remaining in proximity (i.e., stay) and some form of social behavior between buddies (i.e., play or talk). The decision not to require *play* and *talk* was that definitions for these behaviors often overlapped and were different across studies (e.g., in one study, *talk* may have included play narration, while in another study, authors considered that a *play* behavior). Experimental designs included single case design studies with at least three opportunities to demonstrate an effect such that a functional relation could be identified (e.g., three potential demonstrations) or group design studies with random assignment of participants to condition (no eligible group design studies were identified). Focal participants were identified based on author report as children who had low social competence who were

potentially benefiting from the intervention. Peer participants were also identified based on author report; they were children trained to interact with the focal participant as a “buddy.”

Search Procedures

All identification procedures were conducted by two independent assessors (i.e., interobserver agreement assessed for 100% of steps and potential sources). The primary assessor was a doctoral-level expert in special education and two secondary assessors were master’s students in education programs, both of whom were completing coursework and fieldwork to become certified behavior analysts. 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*” using the PsycINFO and ProQuest Dissertations and Theses databases. We first identified all studies that met all inclusion criteria except the final one listed above (e.g., we identified studies that included non-eligible designs, such as A–B designs). From this search, 16 sources were identified, 2 of which were unpublished duplicates of published studies ($n = 14$ unique sources; 99.7% agreement). Forward and backward searches from the original 14 sources 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, given the nature of the review, 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). Two additional articles were identified (99.6% agreement). Finally, a search was conducted using Google Scholar for sources published “since 2018”. Each term listed above was searched separately (with all three “stay play talk” terms resulting in identical results); one additional unpublished dissertation was identified (99.2% agreement). In sum, 17 unique sources were identified (see screening diagram; Authors, 2019). Of these sources, 7 were excluded because they used non-experimental single case designs and 1 was excluded because it contained only SPT versus SPT comparisons (e.g., SPT with and without certain treatment components; for a list of excluded SPT studies, see Authors, 2019).

Descriptive Coding

Descriptive data were coded in the following domains: 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. Average agreement was 96.7%; see [Authors \(2019\)](#) 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 functional relation (i.e., potential functional relations). 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, and recorded descriptive information about pre-intervention language, social-emotional, and play skills when those were provided 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., 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 non-participant); or blind raters.

Rigor Coding

We evaluated rigor using the Single Case Analysis and Review Framework (SCARF; Ledford, Lane, Zimmerman, Chazin, & Ayres, 2016), which allowed for an assessment of outcomes in relation to rigor. SCARF requires yes/no coding for 44 codes in 7 quality/rigor domains and 3 outcome domains (primary, generalization, maintenance). It results in a quality/rigor score 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. Coding was conducted by two independent assessors, with overall agreement of 94.3% (additional information available via Authors, 2019).

Visual Analysis of Outcomes

Selecting outcomes. Visual analysis was conducted by two independent coders for all eligible outcomes in each design in all sources (i.e., was conducted at the design level). Some studies included multiple potential functional relations for each participant (e.g., 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 9 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; What Works Clearinghouse, 2017). We evaluated the presence of functional relations on a 0 – 4 scale, consistent with typical SCARF procedures. A score of 0 indicated no behavior change; scores of 1–2 indicate inconsistent behavior change with one or more non-effects, and scores of 3–4 indicate consistent behavior change with no non-effects. Generally, scores of 3 and 4 are consistent with determinations that a functional relation exists.

Effect Sizes and Meta-Analysis

Selecting outcomes. For purposes of effect size calculation, we selected one dependent variable for each participant in each source. For peer participants, most researchers measured a behavior we categorized as a “broad social interaction”; thus, we used this variable to estimate peer effects (for Milam, the only study without a variable categorized as a broad social interaction, we used “strategy use” to estimate effects). For focal participants, most studies reported only one outcome measure. For studies that measured more than one, we used the variable that was categorized as a broad social interaction (other measured variables included use of SPT strategies, joint attention, and more narrowly defined social interactions).

Log response ratios (LRR). The LRR effect size is the natural log of proportionate change in the mean level between conditions. This metric is appropriate for directly observed behaviors, as commonly measured in single case designs. Further, 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; 2019a). Finally, 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 relevant behaviors were social behaviors expected to improve.

For the purposes of calculating effect sizes, a graduate student extracted data from graphs using PlotDigitizer (2015). Obtained values were corrected under two conditions. First, if the value was not possible given the range of values (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 min) were not corrected.

We used the SingleCaseES package (Pustejovsky & Swan, 2018) in the R statistical environment (R Core Team, 2018) to calculate LRR estimates. We calculated LRR for each focal participant; in most cases, the estimate was from a single A–B comparison in a multiple baseline or 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 synthesized the LRR effect sizes using two types of multi-level random effects meta-analysis models. First, we estimated separate, univariate models for focal child and for peer implementer behaviors. Second, we estimated a multi-variate model for the joint distribution of focal child and peer implementer effects. All multi-level models included random errors at the level of the study and at the level of the participant. The standard deviation of the study-level random errors describes the degree of heterogeneity in effect sizes across studies, as might arise due to variation in implementation, outcomes, settings, or participant inclusion criteria; the standard deviation of the participant-level random errors describes the extent of effect heterogeneity across participants within a study. The multi-variate 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 un-modeled auto-correlation (Pustejovsky, 2018).

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 multi-level random effects model. At each level, larger SDs 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 (PI; Borenstein Higgins, Hedges, & Rothstein, 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, 2019b) packages in the R statistical environment (R Core Team, 2018) for estimating the meta-analysis models and obtaining cluster-robust variances, respectively. Raw data and the R script for replicating the meta-analysis are available in Supplemental Materials (Authors, 2019).

Results

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 (Goldstein, English, Schafer, & Kaczmarek, 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. Researchers in seven of nine sources measured both focal participant 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).

Descriptive Analysis

Participants. There were 25 focal and 31 individual peer participants. 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 8 of 9 sources) and 59 months for peers (range = 36–101 months; reported in 5 of 9 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, 7 focal participants and 7 peer implementers were identified as White, 3 focal participants and 5 peer implementers were identified as Black, and 1 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 ($n = 11$). Others 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 had disabilities; authors reported they were chosen as implementers because they were a sibling with higher social skills than the focal participant or to have a same-gender pairing.

Inclusion criteria. Inclusion criteria for focal participants were unreported in 5 of 9

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 ($n = 1$ each). Of the 8 sources that included individual peer participants, inclusion criteria for implementing peers varied widely and was unreported in 2 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 ($n = 2$), compliance ($n = 2$), 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 2 conducted in the home with siblings, and 1 conducted in a clinic. Of the six sources reporting classroom settings, four were in preschool and two were in elementary school settings. Most sources ($n = 6$) reported that measurement occurred in free play or centers; 2 reported that measurement occurred free play plus other activities (playground, snack, large group), and one reported that measurement occurred during other activities (lunch, PE, art). In one study, the classroom teacher and researchers both facilitated training while in 7 studies researchers facilitated training (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 1–6 different dependent variables, for a total of 25 different variables across studies, in 35 different experimental designs. About half ($n = 13$) were emitted by focal participants, while others ($n = 12$) were emitted by peer participants or by both participants and non-participants (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: (1) *training*, during which children were taught to engage in SPT behaviors, and (2) *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, and talk behaviors was provided. Most ($n = 7$) explicitly reported using role play or practice opportunities, while 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 9 measurement sessions per intervention condition (average range across sources = 5 – 16), for a total average duration of

107 minutes. Presence of prompting and praise were not well-described across sources; authors in 7 sources explicitly reported prompting peer implementers, while presence of prompting was unclear in 2 sources. Authors for only one source explicitly reported prompting of focal participants and one explicitly did not prompt; remaining 7 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 ($n = 6$), while only 2 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 3 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 0–4; actual range 1.0 – 3.2); see scatterplots (Authors, 2019). 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 fewer than 1/3 of designs including measurement of maintained behavior change, and social validity (primarily resulting from low scores for all designs from Severini, Ledford, Barton, & Osborne, 2019, a study in which social validity was not assessed).

Outcomes

Visual analysis. Visual analysis was conducted for 35 designs (from 9 sources). For focal participants, data in 6 of 15 designs (40%) demonstrated a functional relation (scores of 3 or 4); lack of functional relation was primarily due to inconsistent effects (e.g., one non-responder in a multiple baseline across participants design) or failure 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. As shown in the SCARF scatterplots (Authors, 2019), for focal participants, six designs had high quality/rigor scores and consistent positive outcomes while six designs (from four different sources) had high quality/rigor scores and inconsistent or equivocal outcomes. For peer implementers, 11 designs received high rigor and quality ratings and had consistent positive outcomes while 6 designs (5 from a single source; Severini et al., 2019) had high quality/rigor scores and equivocal outcomes. Two sources included only data for focal participants—both demonstrating functional relations. For sources including data for both participant types, two demonstrated functional relations for all peer implementer and focal participant outcomes, one failed to demonstrate a functional relation for either, one demonstrated a functional relation for peer implementer outcomes but not focal participant, and three had mixed outcomes (e.g., demonstrated functional relation for some but not all peer implementer outcomes).

Effect size estimates. LRR effect size estimates are reported in Table 2. 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 all three 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 did not have identified disabilities, but were instead identified as being at risk for social-emotional delays. Given these substantive and methodological differences, we excluded the results of this study from the meta-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; Authors, 2019) 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.

Meta-analysis. Because results from the multivariate and univariate analyses were similar, we report the results from the simpler univariate model in Table 3 (with and without excluding Severini et al., 2019); results from multivariate models are reported in Supplemental Materials (Authors, 2019). 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 3, estimates are similar for the model excluding Severini et al. (average $LRR_i = 1.34$, 282% increase). The 67% prediction interval estimates for the latter model are (156%, 471%) at both the study level and the participant level, which indicate 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. (ES = 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 prediction interval (67% PI = [100%, 558%]), but little within-study variation in effects (estimated participant-level SD = 0.04; both from the model excluding Severini et al.).

Discussion

The purpose of this review was to evaluate the existing evidence regarding the use and outcomes for SPT interventions, which were first assessed more than 20 years ago. 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; therefore, information about the impact of SPT in elementary school settings is incomplete. In terms of visual analysis, a greater proportion of designs demonstrated a functional relation for implementing peers compared to 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 no or less training, (b) the relatively brief intervention period (e.g., about 3 hours, including both training and measurement sessions); and (c) that 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. No study including multiple dependent variables for individual participants demonstrated functional relations for all. This confirms the importance

of identifying which analyses are confirmatory and exploratory, 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 these procedural decisions were rarely reported.

Results of the meta-analysis also indicated variation in the effectiveness of SPT—particularly variation across studies—for both peer and focal participant behavior. Although the estimated average effects were clearly positive, the degree of positive behavior change was less predictable. Some variability in implementation (e.g., training peers only versus buddy groups together) and measurement (e.g., counting the number of occurrences versus 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 improved, thorough reporting practices, it may be possible to explore the relations via meta-regression at a later date.

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 3 of 4 opportunities to do so (Milam, 2018; Severini et al., 2019). Because these two studies were excluded from the meta-analysis altogether (Milam) or in one model (Severini), the reported outcomes primarily represent SPT implemented as a PM intervention rather than a combined or whole-class intervention. Additional research is needed to determine whether there are benefits or detrimental effects when focal children are included in small group training. 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.

After excluding studies without three potential demonstrations of effect, nearly all remaining sources met high standards for single case design; this is discrepant from other bodies of work assessed using SCARF, even those including only peer-reviewed studies (Zimmerman & Ledford, 2017; Zimmerman et al., 2018). 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). Authors of systematic reviews including single case design studies should search for and screen both published studies and grey literature to ensure all relevant evidence is utilized for drawing conclusions.

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. However, data from this review suggest that minimal progress has been made on this front. Relatedly, few studies reported fidelity of peer implementation, a common problem in PM interventions (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 what children might benefit from this intervention. 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, Carter, & Gilson, 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 compare variations of SPT. Specifically, additional studies are 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, (d) training group size, and (e) training group composition. Additional research is also needed to establish the effects of SPT when indigenous facilitators provide the training and the extent to which they can and will implement the procedures. Data are also needed to confirm whether peer implementers maintain behavior change over time, given the lack of maintenance data and relatively short intervention durations reported. Finally, data are also needed to determine whether outcomes for focal children maintain or continue to improve, in the presence or absence of peer implementation maintenance.

Conclusion

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.

References

- Authors (2019). Stay-play-talk meta-analysis: Supplementary materials.
https://osf.io/m263p/?view_only=a0e147cf9b6e4af5a48ad98f9c7fe2c9
- *Baldwin, B. (2013). Peer-mediated intervention and technology: Mobile device application use in implementation of peer buddy training. Unpublished dissertation. Florida State.
- *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., Sigafos, 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. Arlington, VA: Author. Retrieved on April 4, 2019: <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.
- *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*. New York: Routledge.
- *Milam, M. E. (2018). Stay-play-talk with preschoolers: Programming for generalization. Unpublished dissertation. Vanderbilt University.
- 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). Baltimore, MD, US: 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.
- Odom, S. L., & Strain, P. S. (1986). A comparison of peer-initiation and teacher-antecedent interventions for promoting reciprocal social interaction of autistic preschoolers. *Journal of Applied Behavior Analysis*, 19, 59-71.
- 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*.
- 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). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, 24, 217-235.
- Pustejovsky, J. E. (2019b). 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. & 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>
- *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*. doi:
- 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. Retrieved from <http://ocrdata.ed.gov/Downloads/CRDC-School-Discipline-Snapshot.pdf>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. URL: <http://www.jstatsoft.org/v36/i03/>
- 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 (2017). *Procedures and Standards Handbook. Version 4.0*. Retrieved from: <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.

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)	# 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	--
	MB→P	Broad SI	Count	No FR (1) ^a	3
English	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	Percent	FR (4)	1
	A-B-A-B	Specific SI	Percent	FR (4)	1
Milam	AP→P	Play	Duration	FR (4)	--
	AP→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	AP→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. MB = multiple baseline. SP = social partner. P = participant. SI = social interactions. MTS = momentary time sampling. FR = functional relation. # of effect sizes = number of effect sizes calculated from this design (see Table 2). a = low-rigor study. b = not enough information provided to determine specific interval system used. c = For four designs in Severini, data from two peers was combined into a single effect size; two separate functional relation determinations were made.

Table 2. *LRR effect size estimates.*

First author	Buddy Group	Peer Implementer		Focal Participant	
		Effect Size (standard error)	Included Phase Comparison	Effect Size (standard error)	Included Phase Comparison
Baldwin	1	--	--	2.59 (0.35)	3 A–B, 1 P
Barber	1	1.33 (0.29)	1 A–B, 1 P	1.02 (0.23)	1 A–B, 1 P
	2	1.10 (0.25)	1 A–B, 1 P	1.02 (0.15)	1 A–B, 1 P
	3	1.33 (0.22)	1 A–B, 1 P	1.45 (0.18)	1 A–B, 1 P
English	1	1.74 (0.40)	1 A–B, multiple Ps	1.80 (0.26)	1 A–B, 1 P
	2	2.03 (0.44)	1 A–B, multiple Ps	1.77 (0.57)	1 A–B, 1 P
	3	1.63 (0.35)	1 A–B, multiple Ps	1.47 (0.30)	1 A–B, 1 P
	4	2.37 (0.48)	1 A–B, multiple Ps	1.95 (0.35)	1 A–B, 1 P
Goldstein	1	1.24 (0.31)	1 A–B, multiple Ps	1.09 (0.39)	1 A–B, 1 P
	2	1.63 (0.23)	1 A–B, multiple Ps	1.22 (0.34)	1 A–B, 1 P
	3	1.60 (0.22)	1 A–B, multiple Ps	1.48 (0.23)	1 A–B, 1 P
	4	1.92 (0.22)	1 A–B, multiple Ps	1.78 (0.34)	1 A–B, 1 P
Kim	1	1.16 (0.44)	1 A–B, 1 P	0.82 (0.50)	1 A–B, 1 P
	2	0.88 (0.27)	1 A–B, 1 P	0.59 (0.32)	1 A–B, 1 P
	3	0.82 (0.17)	1 A–B, 1 P	0.60 (0.12)	1 A–B, 1 P
Laushey	1	--	--	1.19 (0.33)	2 A–B, 1 P
	2	--	--	0.87 (0.15)	2 A–B, 1 P
Milam	1	6.99 (0.50)	1 A–B, 2 P	1.27 (0.56)	1 A–B, 1 P
	2	6.80 (0.24)	1 A–B, 2 P	0.79 (0.29)	1 A–B, 1 P
	3	5.05 (0.50)	1 A–B, 2 P	0.44 (0.29)	1 A–B, 1 P
Severini	1	2.01 (0.40)	2 A–B, 2 P	-2.63 (0.76)	2 A–B, 1 P
	2	0.45 (0.21)	2 A–B, 1 P	0.02 (0.18)	2 A–B, 1 P
Tsao	2	0.52 (0.46)	1 A–B, 1 P	0.81 (0.39)	1 A–B, 1 P
	3	0.95 (0.29)	1 A–B, 1 P	0.89 (0.34)	1 A–B, 1 P
	4	1.71 (0.38)	1 A–B, 1 P	0.92 (0.37)	1 A–B, 1 P

Note: -- indicates no outcome was reported for this participant. A–B = a single baseline versus SPT comparison. Multiple Ps = metric represents data from multiple peers; author graphed data from multiple buddies as a single data path. 1 P = metric represents data from a single participant. 2 P = metric represents data from 2 participants, 2 peers were paired with one focal child, data for each were graphed separately, and we combined data in a single metric.

Table 3. Results of univariate multi-level meta-analyses

Outcome	Average Effect (SE)	95% CI	LRR		Average Effect	Percentage Change		
			Study-level SD	Participant- level SD		95% CI	67% PI (study)	67% PI (participant)
<i>Including all studies with children with disabilities as focal participants</i>								
Peer Participants	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)
<i>Excluding Severini et al., 2019</i>								
Peer Participants	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 due to inclusion of focal participants without disabilities and near-zero baseline means.