Concordance of multiple informant assessment of school-based social skills intervention and association with child outcomes: Results from a randomized trial

Daina M. Tagavi[1] (ID) **, Kaitlyn Ahlers[2,3]** (ID) **, Alice Bravo[3,4],
Alana J. McVey[3] and Jill Locke[3]** (ID)

Abstract

Background

Fidelity, or the degree to which an intervention is implemented as designed, is essential for effective implementation. There has been a growing emphasis on assessing fidelity of evidence-based practices for autistic children in schools. Fidelity measurement should be multidimensional and focus on core intervention components and assess their link with program outcomes. This study evaluated the relation between intervention fidelity ratings from multiple sources, tested the relation between fidelity ratings and child outcomes, and determined the relations between core intervention components and child outcomes in a study of an evidence-based psychosocial intervention designed to promote inclusion of autistic children at school, Remaking Recess.

Method

This study extends from a larger randomized controlled trial examining the effect of implementation support on Remaking Recess fidelity and child outcomes. Schools were randomized to receive the intervention or the intervention plus implementation support. Observers, intervention coaches, and school personnel completed fidelity measures to rate completion and quality of intervention delivery. A measure of peer engagement served as the child outcome. Pearson correlation coefficients were calculated to determine concordance between raters. Two sets of hierarchical linear models were conducted using fidelity indices as predictors of peer engagement.

Results

Coach- and self-rated completion and quality scores, observer- and self-rated quality scores, and observer- and coach-rated quality fidelity scores were significantly correlated. Higher observer-rated completion and quality fidelity scores were predictors of higher peer engagement scores. No single intervention component emerged as a significant predictor of peer engagement.

Conclusions

This study demonstrates the importance of using a multidimensional approach for measuring fidelity, testing the link between fidelity and child outcomes, and examining how core intervention components may be associated with

[1]Department of Psychology, University of Washington, Seattle, WA, USA
[2]Department of Psychiatry, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA
[3]Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, WA, USA
[4]School of Education, Seattle Pacific University, Seattle, WA, USA

Corresponding author:
Daina M. Tagavi, Department of Psychology, University of Washington, Box 351525, Guthrie Hall, Seattle, WA 98195, USA.
Email: dtagavi@uw.edu

child outcomes. Future research should clarify how to improve multi-informant reports to provide "good enough" ratings of fidelity that provide meaningful information about outcomes in community settings.

Plain Language Summary: Fidelity is defined as how closely an intervention is administered in the way the creators intended. Fidelity is important because it allows researchers to determine what exactly is leading to changes. In recent years, there has been an interest in examining fidelity of interventions for autistic children who receive services in school. This study looked at the relationship between fidelity ratings from multiple individuals, the relationship between fidelity and child outcomes, and the relationship between individual intervention component and child changes in a study of Remaking Recess, an intervention for autistic children at school. Schools were randomly selected to receive the intervention only or the intervention plus implementation support from the research team. Observers, intervention coaches, and individuals delivering the intervention themselves completed fidelity measures. Child engagement with peers was measured before and after the intervention. Several measures of self-, coach-, and observer-report fidelity were associated with each other. Higher observer-reported fidelity was associated with higher child peer engagement scores. No single intervention step was linked to child peer engagement and both treatment groups had similar outcomes in terms of fidelity. This study shows the importance of having multiple raters assess different parts of intervention fidelity, looking at the link between fidelity and child outcomes, and seeing how individual intervention steps may be related to outcomes. Future research should aim to find out which types of fidelity ratings are "good enough" to lead to positive changes following treatment so that those aspects can be used and targeted in the future.

Introduction

Fidelity, or the degree to which an intervention is implemented as designed (Proctor et al., 2011), is commonly viewed as essential to effective implementation (Edmunds et al., 2022). In routine care settings, fidelity is an important measure of intervention implementation success (McLeod et al., 2013; Stirman, 2020), as researchers have demonstrated that fidelity is a crucial antecedent to desired outcomes (Durlak & DuPre, 2008; Schultes et al., 2015). Fidelity can help to explain variation in outcomes, such that understanding providers' fidelity to implementation is critical to determining whether outcomes can be attributed to the particular intervention or to factors related to implementation (Schoenwald & Garland, 2013; Schultes et al., 2015). Thus, there is increased attention on not only intervention outcomes but also the level to which programs are implemented as intended to determine if *how* providers use an intervention impacts overall outcomes.

Given the established importance of fidelity, there has been a growing emphasis on the more consistent assessment of fidelity in "real-world" settings, such as community agencies and schools (Bond & Drake, 2020; Sanetti et al., 2020). However, measuring fidelity is not always straightforward or feasible in these settings due to a lack of agreement on the best and most pragmatic methods for assessing intervention fidelity (Harn et al., 2017; Schoenwald et al., 2011; Sutherland & McLeod, 2022). Current best practice in fidelity measurement is direct observation (Sanetti et al., 2020), specifically the use of observational coding systems with trained raters

(Schoenwald et al., 2011). There are several practical concerns, however, to using observation to measure fidelity, especially in schools, such as time, cost, and level of training required (Sanetti et al., 2021; Schoenwald et al., 2011). Alternatives to observer-rated fidelity include self-report or participant-rated measures, though researchers have raised significant concerns regarding informant accuracy (e.g., social desirability, lack of knowledge regarding the core components of the intervention; Schoenwald et al., 2011). Despite documented limitations to self-reported fidelity ratings, there are perceived benefits for this type of measurement for evidence-based practice (EBP) implementation in schools, including the opportunity to serve as a "self-check" of EBP implementation and the potential to inform supervisory meetings in which EBP feedback is delivered (Hogue, 2022).

While fidelity measurement is reported to vary based on rater, a few studies have used multiple raters and compared ratings, a practice that has the potential to enhance the validity of fidelity measurement via triangulation of ratings (Bond & Drake, 2020; Lillehoj et al., 2004; Schultes et al., 2015). Hansen et al. (2014) compared teacher and observer fidelity ratings for a middle school based EBP and found that teachers rated their fidelity higher than did observers. Brookman-Frazee et al. (2021) examined therapist and observer fidelity ratings of EBP delivery in children's mental health services and found greater concordance between self- and observer- rated fidelity when an EBP had greater structure. Dickson and Suhrheinrich (2021) compared observer, supervisor, and provider fidelity ratings for the delivery of an EBP for autistic youth and found that supervisors had greater concordance with

observers than did providers. Despite these findings, questions remain as to how to pragmatically measure fidelity in school-based research and the association between EBP fidelity and student outcomes.

Moreover, while it may seem as though "more is better" when it comes to fidelity, instead, researchers need to identify the intervention components associated with desired outcomes to determine which components must be delivered to fidelity and which can be adapted to fit the context (Durlak & DuPre, 2008; Sutherland & McLeod, 2022). As a result, research has shifted to identifying the essential, or core, components of an intervention, or those posited to affect outcomes, and optimizing fidelity measures to capture these components while allowing for an "adaptable periphery" to improve fit-to-context (Edmunds et al., 2022). Therefore, it is crucial that fidelity measurement focus on core intervention components and assess their link with program outcomes (Schultes et al., 2015).

Fidelity and Autism-Specific Interventions

There is emerging systematic research examining the implementation of EBPs for autistic children in schools, and the association with student outcomes (e.g., Kratz et al., 2019; Locke, Kang-Yi, et al., 2019; Mandell et al., 2013; Pellecchia et al., 2015; Stahmer et al., 2015; Suhrheinrich et al., 2021). While schools are the primary setting in which autistic youth receive intervention, EBPs are not consistently implemented in schools (Dingfelder & Mandell, 2011). Furthermore, research suggests that intervention fidelity for autism-focused EBPs often is poor (e.g., Pellecchia et al., 2015; Locke et al., 2015, 2019; Mandell et al., 2013; Stahmer et al., 2015); however, students may still demonstrate improved outcomes. For instance, pilot studies and randomized controlled trials support the effectiveness of Remaking Recess, an evidence-based social engagement intervention that promotes the inclusion of autistic children on school playgrounds, in public elementary settings (Kretzmann et al., 2015; Locke, Kang-Yi, et al., 2019; Locke, Shih, et al., 2019; Shih et al., 2019). Studies of Remaking Recess, however, have shown low fidelity, in terms of both quality and adherence, with school personnel using approximately 50% of the intervention components, yet autistic children still demonstrated improved peer engagement (Kretzmann et al., 2015; Locke et al., 2015, 2019, 2020; Shih et al., 2019). These studies collected fidelity reports from multiple sources to assess the consistency between various reports, as well as their predictive ability for youth outcomes. Independent observer-, coach-, and school personnel self-report ratings of fidelity have consistently improved over an intervention training period (Kretzmann et al., 2015; Locke et al., 2015, 2019, 2020) but have declined at 6-week follow-up (Locke et al., 2020; Shih et al., 2019). Interestingly, Locke et al.

(2020) found a higher mean quality fidelity score for self- vs. observer-ratings, but a higher observer- vs. self-rated strategy use fidelity score. Self-rated use also has been found to be higher (90%) than both coach- (76%), and observer-rated (52%) at a post-treatment timepoint, a pattern that maintained at follow-up (self = 85%, coach = 62%, observer = 40%; Shih et al., 2019). These findings are consistent with data from Locke et al. (2019), who found self-rated fidelity to be higher than coach-rated at baseline (self = 54.4%, coach = 38.8%) and at exit (self = 94%, coach = 79.6%). Previous results highlight the importance of capturing fidelity from multiple sources, examining how those ratings do or do not align, and testing how ratings relate to youth social outcomes. Notably, the links between Remaking Recess fidelity overall or to individual core components of the intervention and youth outcomes have not been examined.

Objectives of Current Study

There were three objectives of the current study. (a) *To evaluate the relationship between intervention fidelity ratings provided by multiple sources.* Although findings from the literature were mixed, we expected that fidelity ratings supplied by providers, coaches, and observers would be associated positively with one another. (b) *To examine the relationship between fidelity ratings and youth outcomes.* Based on prior research, we hypothesized that self-, coach-, and observer-rated fidelity would be associated positively with youth social engagement outcomes. (c) *To examine the relationship between each core component of Remaking Recess and youth outcomes.* We hypothesized that quality scores on all components of Remaking Recess would be positively associated with youth outcomes.

Method

The present study was a part of a larger trial examining the impact of implementation support on fidelity of Remaking Recess and child social outcomes (Locke, Shih, et al., 2019). The study was approved by the Institutional Review Boards from the research university, as well as each school district that participated.

Study Design

A randomized controlled design was used with three measurement timepoints: Baseline (start of implementation); Endpoint (post-implementation); and Follow-up (6 weeks post-implementation). Randomization was completed at the school level, with schools randomly assigned to one of two conditions: (a) training in Remaking Recess (*n* = 6), or (b) training in Remaking Recess with implementation support (*n* = 6), as reported in Locke, Shih et al.,

2019. An independent data management core generated a random number sequence to assign randomization.

Procedure

After schools were identified and agreed to participate, research staff distributed recruitment materials to both school personnel and eligible families and met with interested participants to further explain the study and their role as participants. Additionally, research staff met with consented children to explain the study to them in comprehensible language to obtain their assent. Subsequently, school personnel began implementing Remaking Recess based on the condition to which they were randomly assigned. Observers assessed the enrolled children's social engagement on the playground at each timepoint for the duration of the school year. Observers were naïve to study procedures and provider intervention condition.

See Figure 1 for CONSORT Diagram and see Locke, Shih et al., 2019 for complete study protocol.

Intervention

Remaking Recess. Remaking Recess is an evidence-based psychosocial intervention that was designed to support autistic children to engage socially with their peers at school (Kretzmann et al., 2012; Locke et al., 2019, 2020; Shih et al., 2019). A key characteristic of Remaking Recess is that it is designed for sustainability —it is focused on training school personnel who directly support autistic children and their peers day-to-day so that intervention can be maintained once a schools' participation in the research project has ended.

Participating school personnel were individually paired with a "coach" from the research team with whom they met for didactic and interactive training in their natural environment (i.e., at school during lunch). Training sessions lasted

Figure 1
CONSORT Diagram
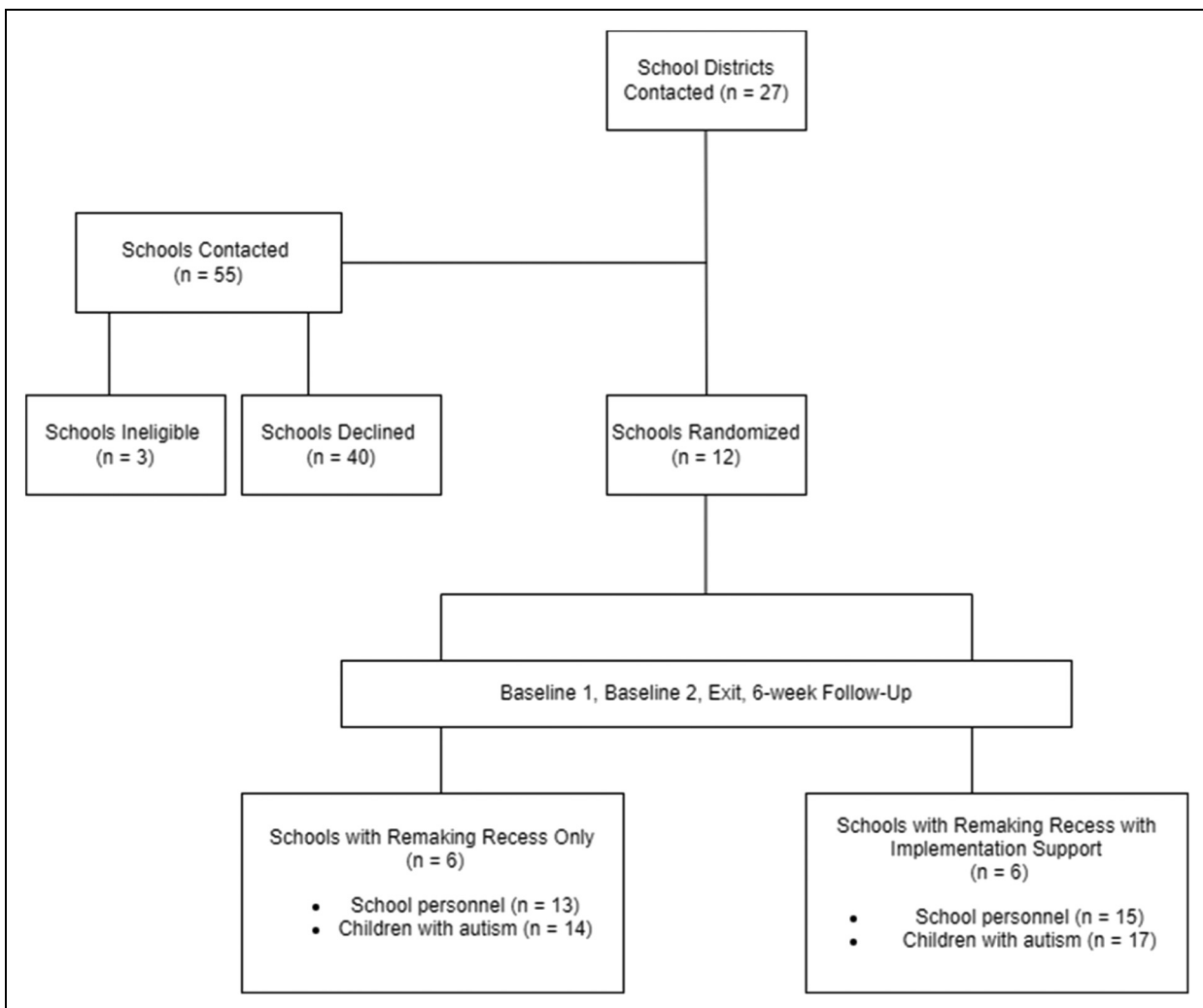
Note. Numbers in bottom most boxes of CONSORT diagram reflect data from the Endpoint included in analyses.

for 30 to 45 minutes and occurred 12 times over 6 weeks. During training sessions, coaches presented information on a target Remaking Recess skill (i.e., what the skill looks like, how it applies to autistic children, and why it is important for a child's development), modeled the skill, and facilitated a practice session where school personnel tried the skill while the coach provided in vivo feedback. School personnel were encouraged to practice skills with students they supported throughout the rest of the week as homework. Enrolled personnel at *all* schools received training in Remaking Recess.

Remaking Recess coaching sessions focused on the following skills: (a) *Attending to child engagement on the playground* (i.e., scanning and circulating the environment for children who may need additional support and identifying children's social engagement states; (b) *Transitioning child to an activity* (i.e., following children's lead, strengths, and interests); (c) *Facilitating an activity* (i.e., providing developmentally- and age-appropriate activities/games to scaffold children's engagement with peers and individualizing the intervention to specific children to help school personnel generalize the intervention to other students in their care); (d) *Fostering communication* (i.e., supporting children's social communicative behaviors, including initiations and responses, and conversations with peers and creating opportunities to facilitate reciprocal social interactions; (e) *Participating in an activity* (i.e., sustain children's engagement within their preferred activities/games and fading out of an activity/gameing to facilitate student independence); (f) *Providing direct instruction of social skills* (i.e., coaching children through difficult situations with peers, should they arise); and (g) *Employing peer models* (i.e., working with peers to engage autistic children; Kretzmann et al., 2015; Locke et al., 2015).

Remaking Recess with Implementation Support. Schools in this condition ($n = 6$) were provided with additional implementation support from MA- or PhD-level coaches. School principals identified school administrators, counselors, psychologists, teachers, and support staff to act as Remaking Recess "champions." Champions did not deliver Remaking Recess, but instead received individualized support from study coaches to address the specific implementation needs of their school. In the first portion of the support process, coaches facilitated the identification of implementation needs across schools (e.g., embed Remaking Recess within the school culture, build internal capacity, improve implementation climate, and provide tangible support and resources). Then, school champions identified the needs that were most pertinent to their school, and coaches worked with champions to develop a plan for implementation based on the identified topic. School champions also were tasked with making announcements about Remaking Recess to faculty, staff, and students, posting visuals of Remaking Recess activities for the week, and facilitating Remaking Recess with other school members/students who were not enrolled in the study.

Coaches. Two coaches from the research team were randomly assigned to provide training in Remaking Recess and Remaking Recess with implementation support for the assigned schools. The coaches were MA- or PhD-level professionals with degrees in education and/or psychology. All coaches were trained in (a) Remaking Recess and (b) the school consultation process, which included working collaboratively with school personnel to deliver an intervention to improve student outcomes (Erchul & Martens, 2010). Coaches were trained by a Remaking Recess developer and evaluated for training and consultation fidelity via a checklist that captured the core intervention components. Coaches were approved to provide training to school personnel once they met $D > 0.80$ fidelity criterion.

Participants

Providers and students were recruited from 12 public elementary schools in five districts in the Northeastern United States over a two-year period. The average school size was 600 students ($SD = 176$), and the average class size was 24.2 students ($SD = 4.5$). On average across schools, 50% of students received free/reduced lunch.

Providers

Twenty-eight school personnel were recruited to implement Remaking Recess with the autistic students they served who were also enrolled in the study. Inclusion criteria for providers included: (a) being employed by the school district (i.e., external/private consultants were excluded), and (b) having at least one autistic student on their caseload with whom they would be willing to work with during the recess period. Twenty-five school personnel served one autistic student each, while three school personnel provided support to two autistic students each. School personnel included 11 teachers and 17 other staff (classroom assistants, one-to-one assistants, and noontime aides). See Table 1 for demographic information for providers and students.

Students

Thirty-one autistic students were recruited to participate in the Remaking Recess intervention. Inclusion criteria for student participants included: (a) an educational classification or medical diagnosis of autism; (b) referral from a school administrator; (c) intelligence quotient scores above 65 as indicated in their school record; and (d) spent at least 80% of their day in a general education elementary setting (kindergarten through fifth grade).

Table 1
Participant Demographics

Child characteristics (n = 31)	M or %
Gender	
Male	90.3%
Female	9.7%
Race	
White	45.1%
Black/African-American	32.3%
American Indian/Alaska Native	0.0%
Asian	9.7%
Hawaiian Native/Pacific Islander	0.0%
Other	3.2%
More than one race	3.2%
None reported	6.5%
Ethnicity	
Hispanic/Latino	12.9%
Grade	
Kindergarten	12.9%
First grade	6.5%
Second grade	16.1%
Third grade	9.7%
Fourth grade	16.1%
Fifth grade	38.7%
School placement	
General education	66.6%
Special education	30.0%
Other	3.3%
Parental education	
Graduate/professional school	41.9%
College graduate	25.8%
Some college	19.4%
High school graduate/GED	9.7%
Some high school	3.2%

Provider characteristics (n = 28)	
Age	39.5
Gender	
Male	14.3%
Female	85.7%
Race	
White	60.7%
Black/African-American	35.7%
Other	3.6%
Ethnicity	
Hispanic/Latino	7.1%
Education	
Graduate/professional school	28.6%
Bachelor's degree	39.3%
Associate's degree	7.1%
High school graduate/GED	25.0%
Autism experience	
Yes	71.4%
No	28.6%
Role	
Teacher	39.3%
Other school personnel	60.7%

Data Collection and Measures

Fidelity Measures

Remaking Recess intervention fidelity was measured from three perspectives: (a) observer-rated, (b) coach-rated, and (c) self-rated. Raters (observer and coach) scored fidelity on seven components of Remaking Recess —attended to child engagement on the playground, transitioned child to an activity, facilitated activity, participated in activity, fostered communication, employed peer models, and provided direct instruction of social skills. *Completion of Remaking Recess* was coded "0" for "No" and "1" for "Yes" based on school personnel's completion of each component. The proportion of completed intervention components divided by the total possible components (seven) was utilized for analyses. *Quality of Remaking Recess delivery*, or how well school personnel used each component of the intervention, was rated on a 5-point Likert scale with "1" indicating "Not Well" to "5" indicating "Very Well" for each component of Remaking Recess that was completed during the intervention session. The average quality rating across all intervention components was utilized for analysis. If a component was not observed, there was no quality rating. School personnel were informed that they would be observed on the playground; however, they were not explicitly informed when data collection would occur or that data were being collected on their use of Remaking Recess steps. This allowed research staff to assess whether the implementation of the intervention naturally occurs without prompts by the research team.

School personnel used a parallel fidelity measure to self-rate their use and quality of intervention delivery. The self-rated form separated the aforementioned seven components into 13 specific sub-components, which mapped directly onto the original components. These were averaged to obtain scores for each of the seven component areas for ease of direct comparison with the observer- and coach-rated fidelity measures. For all raters, fidelity scores at Endpoint were used in analyses. This timepoint was used because it immediately followed school personnel completion of the coaching and consultation process for all Remaking Recess components.

Child Outcome Measure

Playground Observation of Peer Engagement. The Playground Observation of Peer Engagement (POPE; Kasari et al., 2005) is a behavior coding system that examines student social behaviors on a time-interval basis. Independent observers, who were naïve to randomization status, rated students on the playground during recess on the amount of time they spent in solitary play (i.e., unengaged with peers) and joint engagement with peers (i.e., playing turn-taking games, reciprocally engaged in an activity, having a back-and-forth conversation).

Observers watched each student for 40 consecutive seconds and then coded for the next 20 seconds; this process was repeated each minute for the duration of recess (e.g., 10–15 minutes). A POPE developer trained all observers. Reliability was collected on 20% of observation sessions, and observers consistently met acceptable reliability (mean percent agreement = 82%; range = 80%–87%; mean Cohen's κ = 90%; range = 87%–93%). Social engagement was expressed as the percentage (range = 0%–100%) of intervals children spent in joint engagement. Change in joint engagement from Baseline to Endpoint was used in final analyses.

Data Analysis

All study data were managed using Research Electronic Data Capture (REDCap), a secure, web-based application designed to support data maintenance for research studies (Harris et al., 2009). Analyses were conducted in RStudio.

For aim one, to determine the concordance between observer-, coach-, and self-rated fidelity, percent agreement and Pearson correlation coefficients were calculated. For aims two and three, to examine the relation between fidelity and child socialization outcomes, two sets of hierarchical linear models were conducted using fidelity indices as predictors of social engagement scores. All models were analyzed using maximum likelihood estimation to avoid list-wise deletion. All analyses were completed on an intent-to-treat basis. Predictor variables (overall fidelity scores and individual fidelity components) were standardized. We included nesting by school (i.e., children within school) as a random effect in the mixed models. Model assumptions were checked for all final models (i.e., normality of residuals, homoscedastic variances, and appropriate covariance structures). Provider characteristics (i.e., age, gender, ethnicity, education level, years experience working with autistic individuals, and role) were not significantly associated with social communication outcomes and, thus, were not included in the models. Child age and grade were significantly associated

with social communication outcomes and were included as covariates in the models predicting child social engagement (child ethnicity and gender were not associated with outcomes). Provider years experience working with autistic individuals and role were significantly associated with observer-rated fidelity scores. Naïve observers tended to rate teachers Remaking Recess fidelity lower than other school personnel (i.e., paraprofessionals, aides, and counselors).

For the first set of models, we separately examined the association of each fidelity score (observer-, coach-, and self-rated Fidelity Component Completion [hereafter: Completion] and Fidelity Quality of Delivery [hereafter: Quality]) with social engagement. In the second set of models, the seven individual components of fidelity from each of the three raters were simultaneously examined as the independent variables. Intervention status was included as a covariate because the sample included providers in both Remaking Recess and Remaking Recess with implementation support groups. Covariates were held constant to facilitate comparison across models. As such, the results are interpreted as the extent to which the use of Remaking Recess core components relates to student social gains, independent of study group membership.

Results

Intervention Fidelity

Means, standard deviations, and zero-order correlations among all variables are given in Table 2. There was high variability of fidelity scores across raters (Completion *M* range min-to-max = 0.16–0.87; Quality *M* range = 1.58–2.71; see Figures 2 and 3). There also was a high variation in school personnel's fidelity when implementing the prescribed core components across raters (Completion *SD* range = 0.25–0.30; Quality *SD* range = 1.04–2.04). Out of all intervention components, all raters endorsed that school personnel engaged in *Attended to Child Engagement on the Playground*, *Transitioned to an Activity*, and *Set Up an Activity* most consistently. All

Table 2

Mean, SDs, and Zero-Order Disaggregated Correlations for Variables Used in Analysis (n = 31)

Measure	*M*	*SD*	1.		2.		3.		4.		5.		6.		7.
Outcome															
1. Δ Joint engagement	0.21	0.36	-												
Predictors															
2. Self-rated fidelity completion	0.87	0.25	0.23		-										
3. Self-rated fidelity quality	2.71	1.04	0.32		0.71	**	-								
4. Coach-rated fidelity completion	0.75	0.29	0.03		0.50	**	0.47	**	-						
5. Coach-rated fidelity quality	3.60	1.11	0.05		0.57	**	0.49	**	0.83	**	-				
6. Observer-rated fidelity completion	0.16	0.26	0.46	**	0.20		0.41	*	0.34		0.32	*	-		
7. Observer-rated fidelity quality	1.58	2.04	0.46	**	0.16		0.43	*	0.36	*	0.27		0.86	**	-

*p < .05. **p < .01.

raters reported that providers least consistently *Employed Peers to Direct and Redirect Each Other*, while observers and coaches (but not the providers themselves) reported

Figure 2
Implementation Completion Fidelity Scores Across Raters
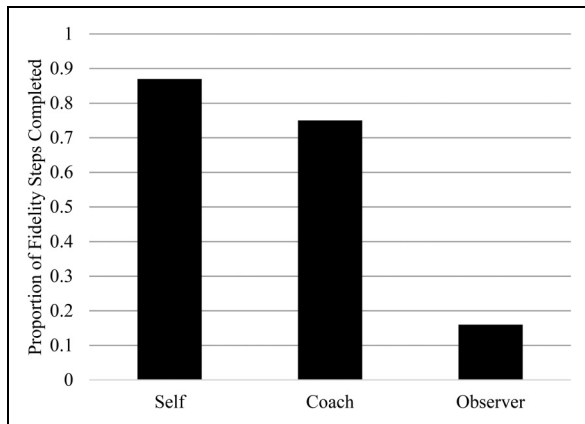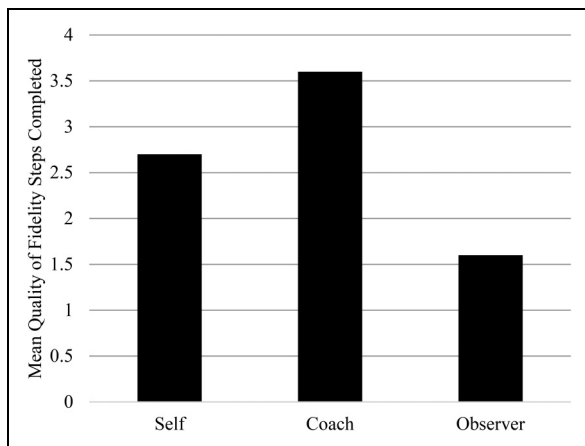

Figure 3
Implementation Fidelity Quality Scores Across Raters

that personnel did not *Foster Communication Between Peers* as consistently as other components.

Percent agreement of fidelity ratings tended to be higher between Self- vs Coach-Ratings and lower between Self- vs Observer- and Coach- vs Observer-Ratings. In term of specific Remaking Recess components, percent agreement between raters tended to be lower overall for fidelity ratings of *Fostering Communication*, *Participating in an Activity*, *Providing Direct Instruction of Social Skills*, and *Employing Peer Models*. See Table 3 for specific proportion of components completed by reporter and agreement on occurrence.

Coach- and self-rated overall Completion and Quality fidelity scores were significantly correlated (Completion: $r = .50$, $p = .004$; Quality: $r = .49$, $p = .005$). Observer- and self-rated Quality scores were significantly correlated ($r = .43$, $p = .02$), but correlations for overall Completion between these two raters were not significant. Observer- and coach-rated Quality scores were significantly correlated ($r = .39$, $p = .03$), but correlations for overall completion between these two raters were not significant. Table 1 depicts all correlations and descriptive statistics of the variables measuring intervention fidelity, as well as the students' outcome variables.

Relation Between Overall Intervention Fidelity and Students' Change in Social Engagement

The intercept-only (or "empty") model was specified to evaluate the intraclass correlations. The mean change in child social engagement score was 0.21 (or 21% spent jointly engaged; Cohen's $d = 0.58$, medium effect; Sullivan & Fein, 2012), which was significantly different from zero. Next, predictors were added to six separate models for each rater and type of fidelity measure. Both observer-rated Completion of Remaking Recess ($b = 0.13$, s.e. $= 0.05$, $p = .02$) and Quality of delivery ($b = 0.15$, s.e. $= 0.05$, $p < .01$) were significant predictors of child

Table 3
Proportion of Components Completed by Reporter and Agreement on Occurrence (n = 31)

Component	Self	Coach	Observer	Self/coach agreement	Self/observer agreement	Coach/observer agreement
Attending to child engagement on the playground	96.8%	96.8%	38.7%	100%	41.9%	41.9%
Transitioning child to an activity	90.3%	83.9%	22.6%	87.1%	32.3%	38.8%
Facilitating an activity	90.3%	74.2%	22.6%	77.4%	32.3%	48.4%
Fostering communication	77.4%	67.7%	6.5%	71.0%	29.0%	38.8%
Participating in an activity	90.3%	71.0%	12.9%	80.6%	22.6%	35.5%
Providing direct instruction of social skills	87.1%	83.9%	6.5%	83.9%	19.4%	22.6%
Employing peer models	77.4%	45.2%	3.2%	48.4%	25.8%	58.1%
Total	87.1%	74.7%	16.1%	78.3%	29.0%	40.6%

social engagement scores. More specifically, higher rates of Completion and higher Quality scores were predictors of higher social engagement scores. Self- and coach-rated Completion and Quality were not significant predictors of child social engagement outcomes (see Table 4 for model results). Intervention condition was not a significant covariate in any of the models (p's > .39). See Figure 4 for predicted joint engagement change scores from fidelity ratings.

Table 4

Multilevel Model Results Predicting Child Social Engagement From Fidelity Ratings Estimation (n = 31 Children From 12 Schools)

	Child social engagement				
	Coefficient	SE	t	df	p
Model 1					
Intercept	1.72	0.49	3.53	30	<.01
Condition	−0.09	0.12	−0.71	30	.48
Child age	−0.29	0.08	−3.44	30	<.01
Child grade	0.32	0.09	3.60	30	<.01
Self-rated fidelity completion	−0.01	0.07	−0.04	30	.97
Model 2					
Intercept	1.63	0.49	3.29	30	<.01
Condition	−0.06	0.13	−0.49	30	.63
Child age	−0.28	0.09	−3.20	30	<.01
Child grade	0.32	0.09	3.35	30	<.01
Self-rated fidelity quality	0.02	0.07	0.33	30	.74
Model 3					
Intercept	1.74	0.43	4.01	30	<.01
Condition	−0.09	0.11	−0.86	30	.39
Child age	−0.29	0.08	−3.69	30	<.01
Child grade	0.32	0.09	3.66	30	<.01
Coach-rated fidelity completion	−0.02	0.06	−0.37	30	.71
Model 4					
Intercept	1.71	0.43	3.98	30	<.01
Condition	−0.09	0.11	−0.78	30	.44
Child age	−0.29	0.08	−3.66	30	<.01
Child grade	0.32	0.09	3.65	30	<.01
Coach-rated fidelity quality	−0.01	0.06	−0.02	30	.98
Model 5					
Intercept	1.45	0.41	3.57	30	<.01
Condition	−0.01	0.10	−0.03	30	.97
Child age	−0.25	0.07	−3.41	30	<.01
Child grade	0.29	0.08	3.54	30	<.01
Observer-rated fidelity completion	0.13	0.05	2.42	30	.02
Model 6					
Intercept	1.52	0.37	4.07	30	<.01
Condition	−0.02	0.10	−0.17	30	.87
Child age	−0.27	0.07	−3.94	30	<.01
Child grade	0.32	0.08	4.17	30	<.01
Observer-rated fidelity quality	0.15	0.05	3.19	30	<.01

Relation Between Core Component Fidelity and Students' Change in Social Engagement

In the second set of models, no single component of *Remaking Recess* (as measured by any of the three raters) emerged as a significant predictor of child social engagement outcomes (see Table 5 for model results).

Discussion

Accurate fidelity measurement is important for understanding intervention outcomes (Schultes et al., 2015), as well as for training users in the adoption of EBPs (Novins et al., 2013). This study evaluated the relationship between multiple sources of fidelity ratings (observer, coach, and self) for a school-based social engagement intervention, Remaking Recess, for autistic students, the relationship between fidelity scores and child outcomes, and the relationship between Remaking Recess' core components and child outcomes.

While there was high variability in fidelity scores across sources, we found positive associations between (a) observer- and coach-ratings of implementation quality, (b) observer- and self-ratings of implementation quality, and (c) coach- and self-ratings of implementation quality and intervention completion. Two key patterns in these findings emerged. First, agreement on specific components was variable, in that there tended to be lower rater agreement for *Fostering Communication, Participating in an Activity, Providing Direct Instruction of Social Skills, and Employing Peer Models,* particularly between observer- and self- and observer- and coach-ratings. One explanation for this may be that in previous studies of Remaking Recess, researchers have posited that *Fostering Communication, Providing Direct Instruction of Social Skills, and Employing Peer Models* are the most difficult components to master (Locke et al., 2019). Another reason for these findings might be that fully participating in play with children also may be more difficult for providers. Additionally, some of these components with lower agreement may also be more challenging for an observer to capture because school personnel have learned to administer them discreetly or before transitioning to recess (e.g., *Providing Direct Instruction of Social Skills, Employing Peer Models*). Finally, this may be explained by the fact that school personnel and their coaches may also have more contextual knowledge about the specific environment or child, which may have led to higher agreements between these raters.

A second pattern that emerged was that provider years working with autistic students was significantly positively associated with observer-ratings of fidelity and provider role was also significantly associated with observer-rated fidelity, in that observers rated teacher fidelity lower overall than other school personnel such as

Figure 4

Predicted Joint Engagement Change Scores From Fidelity Ratings
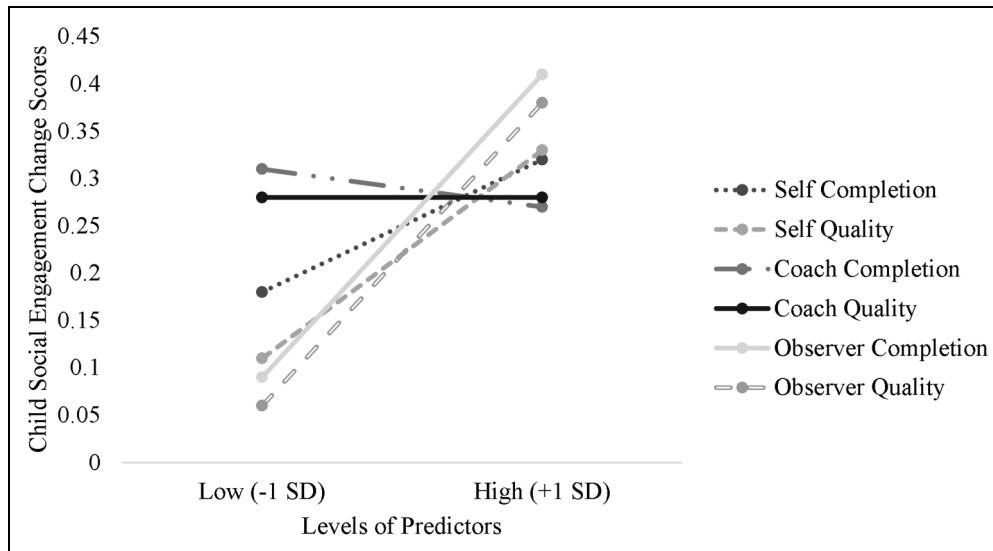
paraprofessionals, aides, and counselors (of note, these characteristics are not associated with child outcomes). While further research is needed to fully understand this pattern, one explanation for this may be that providers with more autism experience or who spend more dedicated time with autistic students are better able to implement Remaking Recess with fidelity. Additionally, differences in teachers' perceptions of their role at recess compared to other professionals' (which may more often involve serving as one-one-one support) could also contribute to differences in fidelity scores. Future Remaking Recess training for broad groups of educators may need to more explicitly address these perceptions.

Only observer ratings of completion and quality significantly predicted child social engagement scores, while no single Remaking Recess component emerged as a significant predictor of child social engagement. Taken together, these findings suggest that it is not one particular component of Remaking Recess driving child outcomes, but instead quality implementation of the packaged intervention as a whole, which may have implications for future training. Findings such as these support the interdependence of components that make up multi-component interventions. Future training efforts should emphasize the coordinated delivery of these components to promote the best chance of high-quality delivery, and thereby optimal outcomes. Our descriptive data also shows a strikingly lower proportion of components completed when rated by observer compared to self and coach, which may be explained by the fact that observers receive more rigorous training in fidelity measurement specifically than educators or coaches (i.e., observers need to achieve 80% reliability on fidelity codes throughout the study). Additionally, observers tend to have higher education levels and

autism expertise. These findings highlight the importance of enhancing training on provider self-assessments to potentially increase the utility of these reports going forward.

Our results strengthen prior identification of observers who are naïve to condition as providing more predictive fidelity data that are less susceptible to bias (Borrelli et al., 2005). Independent observer ratings, however, often are unrealistic in community-based practice settings (Stirman et al., 2017). In school-based implementation research, student privacy concerns frequently preclude video recording as a modality for data collection, yet obtaining live observations of sessions, particularly given the need for inter-rater reliability, is a challenge for even well-funded research studies (Breitenstein et al., 2010; Lillehoj et al., 2004). In addition, live observers are a time- and cost-intensive approach that involves training staff who are typically masked to participant condition (meaning that they cannot help to implement other aspects of the study), reimbursing staff for commuting to and from service settings, and paying staff for their time engaged in observation. COVID-19 concerns have exacerbated challenges with observational fidelity measurement in schools, given the restrictions the pandemic has placed on researchers' abilities to physically enter school grounds. The use of provider (i.e., coach- and self-) fidelity ratings allow for the potential benefits of reduced costs and the promotion of a practice-based learning approach for intervention implementers (Becker-Haimes et al., 2021) and may be more feasible as schools continue to grapple with COVID-19 restrictions and changes to school policies regarding external in-person visitors. Additionally, coach- and self-ratings may provide insight into contextual factors such as individual child presentation or provider style,

Table 5

Multilevel Model Results Predicting Child Social Engagement From Individual Remaking Recess Components Estimation (N = 31 Children From 12 Schools)

	Child social engagement				
	Coefficient	SE	t	df	p
Model 1					
Intercept	0.89	0.53	1.86	29	.10
Condition	0.07	0.14	0.51	29	.61
Child age	−0.16	0.09	−1.68	29	.10
Child grade	0.21	0.09	2.07	29	.05
Self-rated attending	−0.07	0.10	−0.71	29	.49
Self-rated transitioning	−0.16	0.10	−1.58	29	.13
Self-rated facilitating activity	0.16	0.13	1.19	29	.25
Self-rated fostering communication	0.27	0.16	1.66	29	.11
Self-rated participating in activity	−0.19	0.16	−1.17	29	.25
Self-rated in vivo social skills instruction	0.05	0.14	0.33	29	.75
Self-rated employing peer models	0.05	0.06	0.80	29	.43
Model 2					
Intercept	1.73	0.59	2.94	30	<.01
Condition	−0.20	0.12	−1.71	30	.10
Child age	−0.27	0.11	−2.55	30	.02
Child grade	0.28	0.11	2.61	30	.01
Coach-rated attending	0.03	0.11	0.33	30	.74
Coach-rated transitioning	−0.08	0.13	−0.66	30	.61
Coach-rated facilitating activity	0.14	0.08	1.26	30	.22
Coach-rated fostering communication	−0.15	0.09	−1.76	30	.09
Coach-rated participating in activity	−0.07	0.08	−0.87	30	.39
Coach-rated in vivo social skills instruction	0.04	0.11	0.33	30	.74
Coach-rated employing peer models	0.06	0.08	0.72	30	.48
Model 3					
Intercept	1.28	0.37	3.50	30	<.01
Condition	−0.04	0.10	−0.40	30	.69
Child age	−.24	0.07	−3.60	30	<.01
Child grade	0.30	0.08	4.02	30	<.01
Observer-rated attending	−0.06	0.11	0.55	30	.59
Observer-rated transitioning	0.14	0.12	1.57	30	.13
Observer-rated facilitating activity	0.02	0.07	0.14	30	.89
Observer-rated fostering communication	−0.16	0.10	−1.57	30	.13
Observer-rated participating in activity	0.16	0.10	1.64	30	.11
Observer-rated in vivo social skills instruction	0.09	0.10	1.06	30	.29
Observer-rated employing peer models	0.05	0.06	0.71	30	.48

which may be helpful when considering the personalization or individuation of treatment to promote long-term sustainability.

Clearly, a balancing act is needed to use fidelity measurement that provides meaningful information on intervention recipient outcomes, while accounting for privacy, cost of measurement, and sustainability of implementation monitoring systems beyond a research study. These concerns suggest that pragmatic approaches to measurement, such as provider fidelity ratings, are important, particularly for implementation research that is focused on intervention uptake and sustainment (Becker-Haimes et al., 2021). Determining how to train cost-effective and sustainable fidelity monitors to provide "good enough" ratings that

provide meaningful information related to student outcomes is an important next step for implementation research. Regardless of who is completing fidelity ratings, triangulating data across sources is one way to strengthen their validity (Bond & Drake, 2020). Ensuring multiple raters complete fidelity measurement at regular intervals may be an important consideration when relying on self or coach-ratings.

To our knowledge, this is one of a few studies connecting fidelity ratings with targeted child outcomes in school-based research (Lillenhoj et al., 2004). Future research should continue to incorporate fidelity measurement in empirical assessment to better understand how approaches impact accuracy, and how intervention fidelity affects

intervention effectiveness. Additional research is needed to make more definitive conclusions about effective and efficient fidelity measurement systems that can provide meaningful information about recipient outcomes.

While this study provides insight into fidelity measurement across raters, some limitations must be acknowledged. Data were collected across a relatively small participant sample who were geographically located in the Northeastern United States. Future research should include larger, more representative samples from regions across the United States to promote generalizability of findings. Additionally, this study only focused on one intervention for autistic youth in school settings. Further research is needed to establish whether similar patterns occur when measuring fidelity to other interventions for autistic youth in the academic context to determine the universality of these findings and how to best address these issues of fidelity measurement throughout the field of implementation research.

Despite these limitations, this study demonstrated that high levels of observer-rated intervention completion and quality predicted child social engagement outcomes. The findings also suggest that, although observer-ratings were positively associated with each of coach-ratings and self-ratings of quality, coach- and self-ratings were not predictive of child outcomes. Future research should clarify how to improve coach- and self-ratings, perhaps used in combination, to provide "good enough" ratings of intervention completion and quality that provide meaningful information about child outcomes in community settings. Additional work is needed to understand how best to facilitate accurate yet feasible measurement of intervention adherence by self-raters and those who facilitate user implementation.

Clinical Trial Registration

Not applicable.

ORCID iDs

Daina M. Tagavi [ID] https://orcid.org/0000-0002-0521-7474
Kaitlyn Ahlers [ID] https://orcid.org/0000-0001-5786-8140
Jill Locke [ID] https://orcid.org/0000-0003-1445-8509

References

Becker-Haimes, E. M., Klein, M. R., McLeod, B. D., Schoenwald, S. K., Dorsey, S., Hogue, A., Fugo, P. B., Phan, M. L., Hoffacker, C., & Beidas, R. S. (2021). The TPOCS-self-reported Therapist Intervention Fidelity for Youth (TPOCS-SeRTIFY): A case study of pragmatic measure development. *Implementation Research and Practice*. https://doi.org/10.1177/2633489521992553

Beidas, R. S., Maclean, J. C., Fishman, J., Dorsey, S., Schoenwald, S. K., Mandell, D. S., Shea, J. A., McLeod, B. D., French, M. T., Hogue, A., Adams, D. R., Lieberman, A., Becker-Haimes, E. M., & Marcus, S. C. (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: Project FACTS study protocol. *BMC Psychiatry*, *16*(1), 323. https://doi.org/10.1186/s12888-016-1034-z

Bond, G. R., & Drake, R. E. (2020). Assessing the fidelity of evidence-based practices: History and current status of a standardized measurement methodology. *Administration and Policy in Mental Health and Mental Health Services Research*, *47*(6), 874–884. https://doi.org/10.1007/s10488-019-00991-6

Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Czajkowski, S., Breger, R., DeFrancesco, C., Levesque, C., Sharp, D. L., Ogedegbe, G., Resnick, B., & Orwig, D. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology*, *73*(5), 852–860. https://doi.org/10.1037/0022-006X.73.5.852

Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing and Health*, *33*(2), 164–173. https://doi.org/10.1002/nur.20373

Brookman-Frazee, L., Stadnick, N. A., Lind, T., Roesch, S., Terrones, L., Barnett, M. L., Regan, J., Kennedy, C. A., Garland, A. F., & Lau, A. S. (2021). Therapist-observer concordance in ratings of EBP strategy delivery: Challenges and targeted directions in pursuing pragmatic measurement in children's mental health services. *Administration and Policy in Mental Health and Mental Health Services Research*, *48*(1), 155–170. https://doi.org/10.1007/s10488-020-01054-x

Dickson, K. S., & Suhrheinrich, J. (2021). Concordance between community supervisor and provider ratings of fidelity: Examination of multi-level predictors and outcomes. *Journal*

of Child and Family Studies, 30(2), 542–555. https://doi.org/10.1007/s10826-020-01877-0

Dingfelder, H. E., & Mandell, D. S. (2011). Bridging the research-to-practice gap in autism intervention: An application of diffusion of innovation theory. Journal of Autism and Developmental Disorders, 41, 597–609.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. American Journal of Community Psychology, 41(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Edmunds, S. R., Frost, K. M., Sheldrick, R. C., Bravo, A., Straiton, D., Pickard, K., Grim, V., Drahota, A., Kuhn, J., Azad, G., Pomales Ramos, A., Ingersoll, B., Wainer, A., Ibanez, L. V., Stone, W. L., Carter, A., & Broder-Fingert, S. (2022). A method for defining the CORE of a psychosocial intervention to guide adaptation in practice: Reciprocal imitation teaching as a case example. Autism, 26(3), 601–614. https://doi.org/10.1177/13623613211064431

Erchul, W. P., & Martens, B. K. (2010). School consultation: Conceptual and empirical bases of practice (3rd ed.). Springer Science + Business Media. https://doi.org/10.1007/978-1-4419-5747-4

Hansen, W. B., Pankratz, M. M., & Bishop, D. C. (2014). Differences in observers' and teachers' fidelity assessments. The Journal of Primary Prevention, 35(5), 297–308. https://doi.org/10.1007/s10935-014-0351-6

Harn, B. A., Parisi Damico, D., & Stoolmiller, M. (2017). Examining the variation of fidelity across an intervention: Implications for measuring and evaluating student learning. Preventing School Failure: Alternative Education for Children and Youth, 61(4), 289–302. https://doi.org/10.1080/1045988X.2016.1275504

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)--A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics, 42(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hogue, A. (2022). Behavioral intervention fidelity in routine practice: Pragmatism moves to head of the class. School Mental Health, 14(1), 103–109. https://doi.org/10.1007/s12310-021-09488-w

Kasari, C., Rotheram-Fuller, & Locke. J., (2005). The development of the playground observation of peer engagement (POPE) Measure. Unpublished manuscript. University of California Los Angeles.

Kratz, H. E., Stahmer, A., Xie, M., Marcus, S. C., Pellecchia, M., Locke, J., Beidas, R., & Mandell, D. S. (2019). The effect of implementation climate on program fidelity and student outcomes in autism support classrooms. Journal of Consulting and Clinical Psychology, 87(3), 270–281. https://doi.org/10.1037/ccp0000368

Kretzmann, M., Locke, J., & Kasari, C. (2012). Remaking recess: The manual [Unpublished manuscript funded by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS) under grant number UA3 MC 11055 (AIR-B)]. Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles. http://www.remakingrecess.org

Kretzmann, M., Shih, W., & Kasari, C. (2015). Improving peer engagement of children with autism on the school playground:

A randomized controlled trial. Behavior Therapy, 46(1), 20–28. https://doi.org/10.1016/j.beth.2014.03.006

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. Implementation Science, 10(155), 1–17. https://doi.org/10.1186/s13012-015-0342-x

Lillehoj, C. J., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. Health Education and Behavior, 31(2), 242–257. https://doi.org/10.1177/1090198103260514

Locke, J., Kang-Yi, C., Frederick, L., & Mandell, D. S. (2020). Individual and organizational characteristics predicting intervention use for children with autism in schools. Autism, 24(5), 1152–1163. https://doi.org/10.1177/1362361319895923

Locke, J., Kang-Yi, C., Pellecchia, M., & Mandell, D. S. (2019). It's messy but real: A pilot study of the implementation of a social engagement intervention for children with autism in schools. Journal of Research in Special Educational Needs, 19(2), 135–144. https://doi.org/10.1111/1471-3802.12436

Locke, J., Olsen, A., Wideman, R., Downey, M. M., Kretzmann, M., Kasari, C., & Mandell, D. S. (2015). A tangled web: The challenges of implementing an evidence-based social engagement intervention for children with autism in urban public school settings. Behavior Therapy, 46(1), 54–67. https://doi.org/10.1016/j.beth.2014.05.001

Locke, J., Shih, W., Kang-Yi, C. D., Caramanico, J., Shingledecker, T., Gibson, J., Frederick, L., & Mandell, D. S. (2019). The impact of implementation support on the use of a social engagement intervention for children with autism in public schools. Autism, 23(4), 834–845. https://doi.org/10.1177/1362361318787802

Mandell, D. S., Stahmer, A. C., Shin, S., Xie, M., Reisinger, E., & Marcus, S. C. (2013). The role of treatment fidelity on outcomes during a randomized field trial of an autism intervention. Autism, 17(3), 281–295.

McLeod, B. D., Southam-Gerow, M. A., Tully, C. B., Rodríguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. Clinical Psychology: Science and Practice, 20(1), 14–32. https://doi.org/10.1111/cpsp.12020

Mettert, K., Lewis, C., Dorsey, C., Halko, H., & Weiner, B. (2020). Measuring implementation outcomes: An updated systematic review of measures' psychometric properties. Implementation Research and Practice, 1(1), 1–29. https://doi.org/10.1177/2633489520936644

Novins, D. K., Green, A. E., Legha, R. K., & Aarons, G. A. (2013). Dissemination and implementation of evidence-based practices for child and adolescent mental health: A systematic review. Journal of the American Academy of Child and Adolescent Psychiatry, 52(10), 1009–1025.e18. https://doi.org/10.1016/j.jaac.2013.07.012

Pellecchia, M., Connell, J. E., Beidas, R. S., Xie, M., Marcus, S. C., & Mandell, D. S. (2015). Dismantling the active ingredients of an intervention for children with autism. Journal of Autism and Developmental Disorders, 45, 2917–2927.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. Administration and

Policy in Mental Health and Mental Health Services Research, 38, 65–76.

Sanetti, L. M. H., Charbonneau, S., Knight, A., Cochrane, W. S., Kulcyk, M. C., & Kraus, K. E. (2020). Treatment fidelity reporting in intervention outcome studies in the school psychology literature from 2009 to 2016. *Psychology in the Schools*, 57(6), 901–922. https://doi.org/10.1002/pits.22364

Sanetti, L. M., Cook, B. G., & Cook, L. (2021). Treatment fidelity: What it is and why it matters. *Learning Disabilities Research & Practice*, 36(1), 5–11.

Schoenwald, S. K., & Garland, A. F. (2013). A review of treatment adherence measurement methods. *Psychological Assessment*, 25(1), 146–156. https://doi.org/10.1037/a0029715

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health*, 38(1), 32–43. https://doi.org/10.1007/s10488-010-0321-0

Schultes, M. T., Jöstl, G., Finsterwald, M., Schober, B., & Spiel, C. (2015). Measuring intervention fidelity from different perspectives with multiple methods: The reflect program as an example. *Studies in Educational Evaluation*, 47(1), 102–112. https://doi.org/10.1016/j.stueduc.2015.10.001

Shih, W., Dean, M., Kretzmann, M., Locke, J., Senturk, D., Mandell, D. S., Smith, T., Kasari, C., & Kasari, C. (2019). Remaking recess intervention for improving peer interactions at school for children with autism spectrum disorder: Multisite randomized trial. *School Psychology Review*, 48(2), 133–144. https://doi.org/10.17105/SPR-2017-0113.V48-2

Stahmer, A. C., Rieth, S., Lee, E., Reisinger, E. M., Mandell, D. S., & Connell, J. E. (2015). Training teachers to use evidence-based practices for autism: Examining procedural implementation fidelity. *Psychology in the Schools*, 52(2), 181–195.

Stirman, S. W. (2020). Commentary: Challenges and opportunities in the assessment of fidelity and related constructs. *Administration and Policy in Mental Health*, 47(6), 932–934. https://doi.org/10.1007/s10488-020-01069-4

Stirman, S. W., Pontoski, K., Creed, T., Xhezo, R., Evans, A. C., Beck, A. T., & Crits-Christoph, P. (2017). A non-randomized comparison of strategies for consultation in a community-academic training program to implement an evidence-based psychotherapy. *Administration and Policy in Mental Health*, 44(1), 55–66. https://doi.org/10.1007/s10488-015-0700-7

Suhrheinrich, J., Melgarejo, M., Root, B., Aarons, G. A., & Brookman-Frazee, L. (2021). Implementation of school-based services for students with autism: Barriers and facilitators across urban and rural districts and phases of implementation. *Autism*, 25(8), 2291–2304.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, 4(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Sutherland, K. S., & McLeod, B. D. (2022). Advancing the science of integrity measurement in school mental health research. *School Mental Health*, 14(1), 1–6. https://doi.org/10.1007/s12310-021-09468-0