



Published in final edited form as:

School Ment Health. 2009 September 1; 1(3): 118–130. doi:10.1007/s12310-009-9006-9.

Using Participatory Action Research to Design an Intervention Integrity System in the Urban Schools

Rebecca Lakin Gullan, Betsy E. Feinberg, Melanie A. Freedman, Abbas Jawad, and Stephen S. Leff

Department of Psychology, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

Abstract

While integrity is often thought of as the degree to which a program is applied as intended, researchers have recently widened the lens to include not only monitoring of program content, but also evaluating the process by which interventions are implemented and the extent to which the intervention is received as intended. Further, a partnership-based approach has been identified to be as critical to facilitating appropriate and accurate monitoring and interpretation of intervention integrity in the cultural context. Building on these expanded definitions of intervention integrity, this study describes how an intervention monitoring system was developed through participatory research in the context of a classroom-based aggression prevention program for students in an inner-city elementary school. The system highlighted evaluation of the quality of intervention delivery and participant responsiveness. Factor analysis, descriptive statistics, and comparison to a less nuanced integrity monitoring system provided information on the informativeness of this new system. Preliminary investigation, however, suggested that future research is necessary to examine the extent to which differences in quality of implementation across classrooms predict clinically significant differences in program outcomes.

Keywords

Intervention integrity; Prevention program; School-based intervention; Participatory action research; Fidelity

Introduction

In recent years, intervention evaluation research has emphasized the need to evaluate the extent to which programs have been implemented as intended (i.e., program integrity). This expectation suggests that programs implemented at a higher level of integrity will produce the strongest and most consistent findings (Bellg et al., 2004; Moncher & Prinz, 1991). Monitoring how programs are carried out can also provide insight as to components that are critical to intervention success (Perepletchikova & Kazdin, 2005) and the feasibility of implementing the intervention (Peterson & McConnell, 1993). Unfortunately, attention to intervention integrity has often been neglected in favor of evaluating program outcomes (Dusenbury, Brannigan, Falco, & Hansen, 2003). For example, Perepletchikova, Treat, & Kazdin (2007) found that only 3.5% of randomized control trials published in six key psychology and psychiatry journals “adequately addressed” treatment integrity. Even in studies that have examined intervention integrity, very few investigated the relation between program integrity and participant

outcomes. For example, in Dane and Schneider's (1998) review of 162 treatment outcome studies, they found that only 39 (24%) recorded program fidelity and a mere 13 (8%) examined the impact of program integrity on intervention effects.

Defining Intervention Integrity

Integrity is broadly defined as the extent to which an intervention is implemented as intended (Gresham et al., 1993). Treatment integrity encompasses three key issues: therapist adherence (i.e., implementing key intervention components), therapist competence (i.e., ability to implement program effectively), and treatment differentiation (i.e., relative effect of different treatment components) (Perepletchikova & Kazdin, 2005). In addition, the definition of integrity evaluation has been expanded to address not only the extent and quality of intervention implementation, but also participant response (Dane & Schneider, 1998). Participant response relates to the understanding that it is not only the "dose delivered," but also the "dose received" that is critical to programs being fully implemented (Linnan & Steckler, 2002). For example, interventionists might consistently implement 100% of program components (dose delivered), but outcomes may differ based on the extent to which participants are actively involved in the intervention (dose received). Thus, it is critical that programs not only evaluate *what* is implemented, but also *how* it is implemented and *how much* participants are engaged in the process (Waltz, Addis, Koerner, & Jacobson, 1993).

Participatory Action Research

Development of a system to effectively monitor intervention integrity requires not only that key program content and procedures be delineated, but also that the definitions of critical program components reflect the cultural context of the intervention. One way cultural responsiveness of interventions and measurement tools can be maximized is through the use of a participatory action research paradigm (e.g., PAR, Leff, Costigan, & Power, 2004; Nastasi et al., 2000). This methodology involves close collaboration with key community stakeholders such as students, teachers, and community members (Nastasi et al., 2000). Specifically, input from individuals living and working in the target school or neighborhood provide critical insight on how empirically based programs can be optimally effective in the context of the targeted community. The PAR methodology is particularly important for research with under-resourced communities, as it facilitates the development of empirically grounded measures and interventions within the context of community resources and needs (Leff et al., 2006). The present study describes how this partnership-based process was used to develop a system to evaluate integrity for a classroom-based aggression prevention program.

Youth Aggression

According to the National Center for Education Statistics (2007), in the course of one school year, youth ages 5–18 were involved in more than 600,000 violent crimes and over a third of high school students said they had been in a fight on or off school grounds. In response to such alarming statistics, efforts to address youth aggression have been implemented at the individual and group level, targeting children with a history of aggressive behavior, and those at-risk. Increasingly, the importance of providing prevention programming to all youth has also been highlighted as critical to reducing overall levels of child aggression (Eisenbraun, 2007). School-based universal prevention programs have been identified as a particularly effective means to reach all students and to teach the skills necessary to reduce aggressive actions before such behaviors lead to more serious violent acts (Bilchik, 2007; Loeber, Lacourse, & Hornish, 2005). In support of these efforts, a review of 53 school-based prevention programs by the Task Force on Community Preventive Services (2007) found that a range of these programs decreased physical aggression across ages (preschool to high school) and settings (e.g., low SES, high violence neighborhoods).

Relational Aggression

Although acts of physical aggression tend to be the focus of news media and prior intervention efforts (see Leff, Power, Manz, Costigan, & Nabors, 2001), increased attention is being paid to more covert or indirect forms of aggression, also known as social or relational aggression. Relational aggression includes acts such as threatening to withdraw friendship, social exclusion, and spreading rumors, and is more frequently associated with females (Cairns, Cairns, Neckerman, Ferguson, & Gariepy, 1989; Crick & Grotpeter, 1995; Galen & Underwood, 1997). Similar to physical aggression, relational aggression has been found to relate to a range of psychological, social, academic, and behavioral problems (e.g., truancy, depression, anxiety, failing grades; Murray-Close, Ostrov, & Crick, 2007; Woods & Wolke, 2004) and can often be a precursor to physical fights (Talbot, Celinksa, Simpson, & Coe, 2002). Further, the failure to include relationally aggressive acts in the assessment process has the potential to under-identify 80% of aggressive girls and 40% of aggressive boys (Crick & Grotpeter, 1995). Despite research attesting to the harmful affects of relational aggression, efforts to address relational aggression have lagged behind those focused on physical aggression (Leff et al., 2001).

Aggression in Urban Youth

Although physical and relational aggression is common across many settings, minority youth from urban environments tend to experience a higher incidence of physical aggression and violence, often associated with lower socioeconomic status (Eisenbraun, 2007). As such, efforts to address and prevent aggressive behavior are particularly critical for these at-risk youth within urban schools. Establishing effective and sustainable interventions, however, requires that programs be developed and evaluated in the context of community resources and needs (Leff et al., 2004). Indeed, research has found that deviations from traditional manual-based interventions often take place in an effort to make programs more culturally relevant (Dusenbury, Brannigan, Hansen, Walsh, & Falco, 2005). Methodologically strong research calls for these adaptations (or areas of flexibility) to be formally built into the intervention in order to maintain essential components yet maximize the responsiveness to the local school and community (Dusenbury et al., 2005).

The Present Study

The present study describes the design and implementation of a system to monitor intervention integrity in the context of the preventing relational aggression in schools everyday (PRAISE) program. PRAISE is a classroom-based aggression prevention program, designed to target both relational and physical aggressions in urban youth. PRAISE is based upon a social-cognitive and ecological/systems model and teaches urban 3rd to 5th graders social information processing, anger management, empathy awareness, and perspective-taking skills (Leff et al., 2008). The intervention takes place at the classroom level with all students participating as part of the school curriculum. PRAISE is 20 sessions long (40 min per session) and uses cartoons, video illustrations, and role plays that were adapted through partnership for use with African American inner-city youth. Three advanced graduate students serve as co-therapists for each classroom participating in PRAISE. All therapists participated in live observation and weekly supervision with a licensed clinical psychologist. In addition, teachers are encouraged to actively participate in facilitating session delivery, e.g., through eliciting or sharing examples and encouraging students to apply techniques to everyday experiences. PRAISE is based upon one of the first empirically supported relational aggression interventions, the Friend-to-Friend (F2F) program, a 20-session indicated group intervention for relationally aggressive 3rd and 4th grade girls in the urban schools (Leff et al., 2007, in press).

This study utilized a partnership-based methodology to develop and pilot a method of evaluating intervention integrity in the context of the PRAISE program that addressed common

limitations in the literature. After piloting an initial integrity monitoring system targeting program content and process variables (referred to as *System 1* in this article), the decision was subsequently made to develop a second integrity monitoring system to address gaps identified in the original system. Collectively, the two systems evaluated the following aspects of intervention integrity: (a) the extent to which key program components were implemented (*content integrity*), (b) the extent to which facilitators encouraged student participation, utilized appropriate behavior management strategies, demonstrated enthusiasm, and managed time adequately (*process integrity*), and (c) student behavior in the classroom, interest and enthusiasm, and level of distractibility (*dose received*). The goals of the study were the following: (1) to describe the process of working in partnership with key community stakeholders to ensure that intervention integrity was defined in an appropriate, comprehensive, and flexible manner; (2) to describe the resultant systems; (3) to compare information gathered from the two systems, and (4) to preliminarily examine the extent to which participant outcomes related to program integrity.

In regard to Study Goal 3, it was expected that the second integrity monitoring system would relate to key components of the original system, but would also provide a more nuanced picture of the quality of intervention delivery. For Study Goal 4, we hypothesized that classrooms with the highest integrity ratings would also demonstrate greater change scores on key outcome variables targeted by the PRAISE program (see Leff et al., 2008 for full description of study and findings). Specifically, it was predicted that classrooms with greater intervention integrity would be associated with a gain in knowledge related to critical problem-solving steps and a reduction in student- and teacher-reported aggression and hostile attribution bias (tendency to perceive ambiguous stimuli or acts as having hostile intent; Dodge, 1980).

Methods

Participants

Participants included 3rd to 5th grade students across two schools taking part in the PRAISE program (described above). In Year 1 (development and piloting of *Integrity Monitoring System 1*), PRAISE was implemented in four 3rd to 5th grade classrooms at a large, inner-city elementary school with a predominantly African American population. In Year 2 (developing piloting of *Integrity Monitoring System 2*), PRAISE was implemented across five 3rd to 4th grade classrooms (143 total students) in a school with the same demographic profile to the previous year's school. Integrity and outcome data were obtained on 107 students (75%) in these classrooms who participated in pre- and post-test evaluation. Year 1 data were used only for measure development, whereas Year 2 data were used for all data analyses in this investigation.

Outcome Measures

Knowledge Measure—The knowledge of social processing and anger management measure (KSPAMM) is a recently created, culturally sensitive measure of children's knowledge of appropriate means of processing social information and managing anger in situations involving peers. The measure is comprised of 15 multiple-choice items. The KSPAMM has been shown to have strong psychometric properties, as item analyses suggest that almost all items discriminate well between more and less knowledgeable individuals, that the test-retest reliability of the measure is strong ($r = .85$), and that the measure appears to be sensitive to treatment changes over time (Leff, Cassano, MacEvoy, & Costigan, 2008).

Children's Social Behavior Questionnaire, Relational, and Physical Aggression subscales—This teacher-report measure consists of three subscales, two of which were used in this study (relational and physical). The psychometric properties of the children's social

behavior questionnaire (CSB) are supported by factor analysis and strong internal consistency ($r > .93$ for all subscales; Crick, 1996). Validation is provided by its moderate correlations with peer reports (r 's = .57–.79) of physical and relational aggression (Crick, 1996).

Cartoon-Based Hostile Attributional Bias in Relationally Provocative Situations

—This measure is a cartoon-based adaptation of a commonly used hostile attributional bias (HAB) measure (Crick, 1995) for urban African American youth. A HAB score is derived for both relationally and instrumentally provocative social situations. This adapted measure has demonstrated strong internal consistency ($\alpha = .81-.83$ for relational and instrumental situations, respectively), test–retest reliability ($r = .86$ for both subscales), and high rates of acceptability in an inner-city, African American student population (Leff et al., 2006).

Procedures to Address Goal 1: Development of Integrity Monitoring Systems

Initial Integrity Monitoring System Development (System 1)—Initially, the research team reviewed the literature on integrity monitoring systems and determined that it would be important to develop a system through PAR that focused upon the key content elements of the treatment (e.g., procedural integrity) combined with important process variables that help to determine how the program may be received (see Power et al., 2005). Based on a literature review related to therapist competence and therapeutic alliance (e.g., Kendall, Chu, Gifford, Hayes, & Nauta, 1998; Waltz et al., 1993), the research team first identified six process-oriented variables that would likely suggest that the treatment was being implemented in a competent and respectful manner, e.g., encouraging all students to participate and utilization of appropriate behavior management strategies (see below for description of all six variables). Each item was operationally defined and assigned a score of 0, 1, or 2, respectively, indicating whether the variables were not implemented, partially implemented, or fully implemented. In addition, the behavior management literature suggests that improved participation in classroom activities often occur when students are on-task and exhibiting relatively low levels of disruptive behaviors (Good & Grouws, 1977).

Feedback from Key Community Stakeholders—Two school employees who were also actively involved in the participating school and community provided ongoing consultation during the development and initial implementation of both the intervention and the integrity monitoring system. One of the community partners was a school secretary who had worked at one of the target schools for over two decades. The other community partner held several part-time roles at two schools within the school district from which the sample was drawn. She served as a home-school liaison coordinator and also as a classroom assistant. These two partners were identified based on principal recommendation and past involvement of each partner in a community-based research carried out by the research team. These two individuals were able to work with the research team to ensure that both intervention content and the integrity system were responsive to the needs of the local community. Specifically, these two individuals were trained in the intervention content along with program therapists, observed a number of intervention sessions, and worked collaboratively with the research team to operationally define important process-oriented intervention implementation variables identified through the literature review and/or to suggest additional implementation variables that the researchers would have otherwise neglected. Feedback from community partners was ongoing and typically took place in weekly meetings with the research team at the school.

The results of this partnership-process allowed for an integrity monitoring system (*System 1*) that included three to four key content areas of each session (as identified by researchers) and six implementation process variables that were jointly identified, defined, and refined by researchers and community partners. The content items varied based on session, whereas the six process items remained the same across all sessions. Process items were the following: (a)

encouraging all students to participate, (b) being responsive to student or teacher questions/comments, (c) facilitators working well together, (d) involvement of classroom teacher, (e) students' behavior in the classroom, and (f) utilization of appropriate behavior management strategies. Finally, researchers and community partners worked together to develop a relatively straightforward rating scale in order to evaluate how well/how much of each core content and process variable occurred on a three-point Likert scale that included 0 = not implemented, 1 = partially implemented, and 2 = fully implemented.

Development of Integrity Monitoring System 2—The following school year, two advanced graduate students were trained to employ the initial integrity monitoring system during live observations across five 3rd and 4th grade classrooms. Although the system was relatively straightforward to use, the integrity monitors indicated that almost all of the process-oriented implementation variables were rated as being fully implemented despite apparent qualitative differences between classrooms at times. This feedback led the research team to design a second implementation rating system (*Integrity Monitoring System 2*) in order to more clearly differentiate the quality of the intervention sessions across a wider range of key process-related variables.

The new implementation system differed in two important ways from the initial one described. First, *System 2* rated quality of implementation across ten variables instead of six, as the integrity monitoring team felt that there were additional variables that could contribute to intervention success. Of the variables retained from *System 1*, several were modified to more accurately reflect key session processes. Second, the team felt that the existing process variables could be scored using a broader scale, so as to capture greater nuance in intervention implementation. Items that were retained from *System 1* were the following: (1) facilitators working well together, and (2) appropriate use of behavior management strategies. Items that were retained from *System 1* but modified were the following: (1) students' behavior and level of distractibility (modified to include inattention during the session), (2) facilitators encouraging students to participate and setting up a successful session context (modified to include communication about rules and respect), and (3) teacher participation and impact on session (modified to include meaningful session contributions, e.g., furthering discussion through providing appropriate examples). Finally, the four new items were the following: (1) students' interest and enthusiasm in the session, (2) appropriate involvement of helpers, (3) enthusiasm of facilitators, (3) time management/pace of session, and (10) global/general impression of the session.

In addition to modifying and/or adding process variables, *System 2* also allowed raters to evaluate each implementation variable on a scale of 1 to 10 with 1 = extremely poor, 5 = at expected level, and 10 = truly outstanding. Items were operationally defined at the expected level of implementation (5), and raters were instructed to increase or decrease ratings relative to the average "anchor." This scale provided our team with a much greater range of possible responses so that we could better differentiate the quality of treatment implementation between sessions. For all integrity ratings, preliminary ratings were made throughout the sessions, but finalized scores were completed once at the end of the entire intervention session. Independent raters were instructed not to change their scores once the session had ended.

Data Analytic Strategies

In order to address the first study goal of presenting and describing a new integrity monitoring system, we evaluated interrater reliability on both measures, calculated mean scores across all items, and conducted a factor analysis of *System 2*.

Interrater Reliability and Selection of Integrity Rating for Data Analyses—

Following 3 months of training and viewing videos to practice integrity coding, the two integrity monitors were randomly assigned PRAISE sessions to observe intervention classrooms. By the end of the intervention, both monitors had rated between five and eight sessions per classroom. Although both monitors observed the same sessions, they did not share or discuss ratings prior to group supervision sessions. Thus, interrater reliability was calculated based on pre-supervision ratings and used to determine the extent to which the independent ratings of the two integrity monitors agreed.

For *System 1*, therapists were considered in accordance if ratings were the same on the 3-point integrity scale. For *System 2*, integrity monitors were considered in accordance if ratings fell within two points of each other on the 10-point integrity scale. For example, if monitor A rated *student interest and enthusiasm in the session* a 5 while therapist B rated the same item a 7, such ratings would be considered in agreement for this particular session. In general, high interrater agreement was expected given the thorough joint training sessions conducted for the monitors. For all data analyses in this study, data from one monitor were randomly selected for each session so that there was a single integrity score associated with each session observed.

Descriptive Statistics—For individual item analysis across both *Systems 1* and *2* (i.e., evaluating the relative extent to which specific variables were implemented across classrooms), mean scores were used. For example, to examine the extent to which *student interest and enthusiasm in the session* was implemented across classrooms, scores on this item were totaled across all observations and divided by the number of observations. In addition, factor scores were calculated for *System 2* to represent mean integrity across all items for each session. Mean integrity across classrooms was calculated by adding factor scores and dividing by the total number of sessions observed.

Factor Analysis—An exploratory factor analysis using principal component extraction method was conducted on the total sample of classroom-based integrity ratings ($n = 33$) to reduce data. The varimax (orthogonal) rotation procedure was utilized first. The Kaiser–Meyer–Olkin test was run to test for adequate sampling, and the Bartlett test of sphericity used to evaluate if substantial correlations exist between the items. Following the recommendations of Kinnear and Gray (2006) and Field (2005), components were selected based on the following criteria: (1) an inspection of the scree plot and (2) having eigenvalues greater than 1.

To address the second study goal—to examine the relation between *Integrity Monitoring Systems 1* and *System 2*—bivariate correlations were calculated between all integrity items across both systems. Finally, to address Study Goal 3, a Kruskal–Wallis non-parametric one-way analysis of variance (ANOVA) was conducted to determine if outcomes differed significantly across classrooms. If so, mean integrity scores were compared to evaluate the potential role of integrity in producing these effects.

Results

Interrater Reliability Data for Integrity Systems

Results confirmed a high level of inter-rater reliability. For *Integrity Monitoring System 1*, coders had adequate inter-rater reliability across content (99%) and process (89%) items. For *Integrity Monitoring System 2*, 96% of the total integrity ratings (i.e., 270/280 ratings) fell within two points across observers. More specifically, four out of the original ten items maintained 100% agreement, another four items maintained 96% agreement, and one item maintained 93% agreement. Interestingly, the *System 2* integrity item *appropriate involvement of student helpers* exhibited the lowest reliability with 86% of the ratings falling within two points. The integrity monitor therapists reported that because the intervention did not provide

specific responsibilities for the student helpers (i.e., passing out worksheets to classmates versus initiating discussions versus participating in a role play, and so on), the monitors evaluated the use of student helpers in a less systematic manner than they did other variables. This may have contributed to the lower rates of agreement. As a result, this variable was dropped from preliminary analyses, producing a final integrity system that included nine items. Data reported below reflect the new 9-item scale of *System 2*.

Descriptive Statistics

Results suggest that most intervention components across both systems were implemented at an expected or satisfactory level. Descriptive statistics for the items of both integrity systems are presented in Table 1. In *System 1*, 94% of the four content area items and 74% of the six process area items were fully implemented. Item-level data indicated that the process item related to *enthusiasm of facilitators* received the highest percentage of fully implemented ratings (94%), whereas only 44% of the observations for *teacher participation and impact on the session* were fully implemented.

In *System 2*, the most frequently observed integrity item score was 5, indicating implementation at the expected level. The integrity item *student interest and enthusiasm in the session* exhibited the highest mean value across sessions and classrooms. An inspection of descriptive statistics revealed variability between as well as within classrooms across the course of the intervention (see Table 1).

Correlational Analyses

Bivariate correlations were conducted to assess the relationship between *System 1* and the newly developed items in *System 2*. As shown in Table 2, correlations ranged from .03 to .82, and the theoretically similar items correlated strongly and positively as expected. For instance, the *System 1* process item *encouraging all students to participate* was significantly and positively associated with the *System 2* item *facilitators encouraging students to participate and setting up a successful session context*, $r = .67, p < .001$, and the *System 1* process item *utilization of appropriate behavior management strategies* exhibited a strong positive relationship to the *System 2* item *student behavior and level of distractibility*, $r = .82, p < .001$. Further, the *System 1* process item *being responsive to student or teacher questions/comments* was significantly and positively associated with only *System 2*'s *student behavior and level of distractibility*, $r = .38, p < .05$, and *global/general impression*, $r = .36, p < .05$. It is also interesting to note that, in contrast to other *System 2* items, which correlated significantly with multiple *System 1* process items, *time management/pace of session* exhibited a significant association to only one process item (e.g., *facilitators working well together*, $r = .38, p < .05$).

The bivariate correlations between the nine *System 2* integrity item scores are also presented in Table 2. In general, the correlations were in the expected direction and represented large effect sizes. For example, *student interest and enthusiasm in the session* was positively correlated with *enthusiasm of facilitators*, $r = .81, p < .001$, and *facilitators encouraging students to participate and setting up a successful session context*, $r = .73, p < .001$. Although statistically significant, *teacher participation and impact on session* exhibited the smallest associations to other *System 2* items, with correlations ranging between $r = .27$ and $.47$.

Factor Analyses

As described above, an exploratory factor analysis was conducted to reduce the data and create factor scores. The Kaiser–Meyer–Olkin value was .83, indicating adequate sampling, and the significant Bartlett test of sphericity, $p < .001$, confirmed that substantial correlations exist between the items. Results suggested a one-factor solution, which accounted for 67% of the total variance in *System 2* integrity item scores. Communalities among the nine items ranged

from .28 to .90. The *teacher participation and impact on session* and *time management/pace of session* subscales exhibited the lowest communalities, .28 and .40, respectively, while the *global/general impression* item exhibited the highest, .90.

Based on these preliminary analyses, factor analysis was re-run following the removal of the *teacher participation and impact on session* and *time management/pace of session* items. These two items were eliminated consistent with the goal of data reduction and because their values were notably lower than all other items in the scale. Results of the final principal axis factoring are presented in Table 3. The Kaiser–Meyer–Olkin value was .87, indicating adequate sampling, and the significant Bartlett test of sphericity, $p < .001$, confirmed that substantial correlations exist among the items. Communalities ranged from .61 to .88, and not surprisingly, the *global/general impression* item exhibited the largest factor loading, .94. Results again supported a one-factor solution, which accounted for approximately 77% of the variance. This sole factor is labeled *global/general impression*. Factor scores were then generated for further analyses. Of note, the analyses were also conducted using oblique rotation, but a similar one-factor solution was produced.

Differential Impact of Program Integrity

A non-parametric test was conducted after a one-way ANOVA indicated assumption violations due to unequal classroom sizes (i.e., Levene's homogeneity of variance). Consistent with the initial ANOVA results, the Kruskal–Wallis test indicated that no significant differences in change scores existed among the five classrooms with the exception of the teacher-report CSB relational, $\chi^2 = 37.79, p < .001$, and overt subscales, $\chi^2 = 21.94, p < .001$. Post hoc comparisons were conducted within an ANOVA framework to ascertain where differences existed. Results revealed that significant group differences on the CSB relational measures existed between Classrooms A and all other classrooms, all $p < .01$. Specifically, students in Classroom A exhibited, on average, greater increases in relational aggression from pre- to post-test. Regarding the CSB overt subscale, Classrooms A and B both differed significantly from Classrooms D and E, $p < .05$. For this outcome, Classrooms A and B had greater increases in overt aggression as compared to the other classrooms. Based on these specific classroom differences, we were further interested in whether integrity factor scores also significantly differed across classrooms (see Table 4), which could contribute to the significant differences in outcome across these classrooms. A series of planned independent samples' t -tests were conducted. Results revealed that Classroom A ($M = 1.09, SD = .82$) exhibited significantly higher integrity factors scores than Classroom C ($M = -1.07, SD = .90$), $t(9) = 4.11, p < .01$, and compared to Classroom E ($M = -.27, SD = .75$), $t(11) = 3.08, p < .05$. Similarly, Classroom B ($M = .49, SD = .46$) exhibited greater integrity factor scores than Classroom E. This difference was also significant, $t(12) = 2.19, p < .05$.

Discussion

This study describes the process of working in partnership with community stakeholders to develop a new method for monitoring program integrity. Preliminary psychometric information provides insight into how a system to evaluate the quality of program implementation performed in the context of a school-based intervention. In addition, comparison between two integrity monitoring systems speaks to the relative informativeness of a more detailed system with a more nuanced response scale. Finally, examination of data across five intervention classrooms provides initial information on the relation between treatment integrity and program outcomes.

Item-level and factor integrity scores across classrooms support the hypothesis that an integrity monitoring system with finer distinctions (*System 2*) has the potential to provide a more nuanced assessment of intervention integrity. Thus, whereas almost all of the variables in

System 1 had a median rating at the maximum level (fully implemented), *System 2* demonstrated a larger range of integrity scores across classrooms and variables. Findings also revealed that intervention integrity (per *System 2*) varied across classrooms.

Surprisingly, however, findings did not support the hypothesis that students in classrooms where the program was implemented with a higher level of process integrity (quality of delivery as measured by *System 2*) would improve the most. Indeed, for overt and relational aggression, it appeared that the opposite might have occurred, with classrooms where the program was implemented with the *lowest* rated process integrity demonstrating the *most* program impact. This finding is in contrast to the expectation that greater quality of program delivery promotes stronger and more consistent effects. These unexpected findings are consistent, however, with Dane and Schneider's (1998) review of treatment integrity studies, which found that only 4 of the 13 studies that actually tested the relation between intervention integrity and outcomes demonstrated positive effects of program exposure or adherence (the rest indicating mixed or no effects), and none found a relation between *quality* of delivery and outcomes. Unfortunately, methodological limitations and the small subset of studies formally testing the effect of program integrity precluded a clear understanding of Dane and Schneider's findings.

Although explanations for this observation in this study remain largely speculative, there are several reasons this might have occurred. First, because all ratings of student overt and relational aggression within each of the five classrooms were completed by the same teacher, it is possible that change scores on these outcomes related to a systematic bias in teacher ratings. Second, it might be that an additional process variable—or an alternate definition of one or more of the variables that were assessed—played a more significant role in impacting outcomes for this type of classroom-based prevention program than the process variables included in *System 2*. For example, although *System 2* originally included an item evaluating *teacher participation*, classroom means on this variable suggested that it did not capture the role of the teacher in a way that predicted greater program effects. Future research, however, might define teacher-related process variables in a way that is significantly related to student outcomes.

Another process variable that might relate to program impact is therapist competence. Although integrity scores and program outcomes in this study did not differ based on which therapy team administered the intervention, interventions with less intensively, or consistently trained therapists might produce differential effects based on therapist competence. Indeed, Dane and Schneider (1998) noted in their review that significant integrity effects appeared to occur more frequently in studies that included objective integrity raters (versus the therapists themselves), perhaps due to therapist's inflated self-ratings. Although this study included objective integrity monitors, it is possible that higher intervention quality was not associated with stronger outcomes in this program because even in the classrooms with the lowest process integrity scores, content was consistently and reliably covered. Thus, it might be that *System 2* was sensitive enough to detect qualitative differences in intervention delivery across classrooms, but these qualitative differences were not large enough to interfere with delivery of program content. In other words, intervention quality or process factors might only be critical to the extent that they impact how much program content is delivered. Future research is necessary to test the predictive utility of content on outcomes and the incremental utility of additional integrity factors, such as teacher engagement.

Contributions to the Literature

This study makes several contributions to the literature. First, it provides a real-world example of how a partnership-based process assisted in the development of a culturally sensitive integrity monitoring system that accounts for theoretically and empirically based critical program components and responds to the unique needs of the target community. PAR methodology has the potential to serve as a bridge between empirically based manualized

interventions and real-world practice (Power et al., 2005). Thus, the use of partnership-based procedures to develop a multi-faceted intervention integrity system can ensure accurate implementation and monitoring of key program components and to identify areas within an intervention that clinicians can modify or adapt to better meet the needs of the specific community.

Being ultimately critical to program success and sustainability, these procedures present unique challenges as well. For example, tension can arise if community partners feel that cultural norms are not recognized or appreciated or when researchers feel that critical program components are challenged. In this study, for instance, there were particular challenges around the rating of the qualitative factor of “child behavior.” Whereas the research team and community partners agreed on the operational definition of different child behaviors (e.g., out of seat, talking out of turn), they differed in relation to the value placed on such behaviors. Thus, when researchers rated loud talking and out of seat behavior low on the integrity scale, feedback from a community partner suggested that the atmosphere was consistent with inner-city classrooms and did not reflect a lack of control that interfered with learning, as was the interpretation of the research team. Resolution of these disagreements involved extensive discussion and articulation of important issues that ultimately resulted in a more precise and nuanced measure of child behavior in the context of an urban school.

Results of this study also provide preliminary information on specific items and response scales that might be maximally informative to understanding key components of intervention integrity. Thus, data analyses suggested that *System 2* be reduced to a 7-item scale. In addition, the 10-point response scale was able to more effectively capture the range of intervention integrity across classrooms. The question remains, however, the extent to which statistically significant differences in quality of delivery according to this integrity monitoring system (e.g., a factor score of 1.09 in Classroom A vs. -1.07 in Classroom C) are clinically significant.

Limitations and Future Directions

Although findings from this study are informative to future research on intervention integrity across diverse settings, there are several limitations. First, students were nested within classrooms. For example, the questionnaire assessing relational and physical aggressions was completed for each individual child by his/her teacher. As such, systematic differences across the five teachers might have occurred. Although outcome data were gathered directly from the child for other variables (e.g., knowledge, hostile attribution bias), change scores on these variables did not differ significantly across classrooms. Intervention integrity was also evaluated at the classroom level, resulting in all students within each of the five classrooms having the same integrity score and precluding a formal test of integrity as a mediator of individual child outcomes.

Another potential limitation of this study is related to the fact that *System 2* was developed in response to a need identified during the early stages of the intervention. As such, *System 2* was not introduced until approximately 4 weeks after the start of the intervention. In addition, schedules did not allow integrity monitors to be present at all intervention sessions, therefore, random assignment of monitors to sessions and classrooms resulted in intervention integrity being evaluated across different sessions for each classroom and for a variable number of times (5–8 observed sessions per classrooms). Finally, the decision to have integrity monitors rate items at the end of each observed session produced potential bias due to the recency effect. Monitors were, however, encouraged to make preliminary ratings and notes throughout the sessions, potentially mitigating the concern that therapist, student, and teacher behavior at the end of the session was disproportionately emphasized. Future use of this system, however, might explore potential differences in ratings near the beginning, middle, and end of the session.

Future research should focus on identifying and testing other potentially important variables related to treatment adherence, quality, and participant response. Studies should also include evaluation of outcomes and integrity factors at the individual child level (e.g., observation of individual child engagement). In addition, potential covariates such as teacher skill should be assessed so as to allow statistical control for classroom effects. Further investigation is also needed to examine the utility of process integrity in predicting change in outcomes over and above delivery of program content. In other words, if content is implemented fully, to what extent does the quality of intervention delivery matter?

Finally, research should identify the optimal response scale for integrity evaluation. Although this study moved from a 3-point response (not implemented, implemented partially, and implemented fully) to a more nuanced 1–10 range, results suggest that evaluating the quality of delivery at this level of detail might be unnecessary. Attempts should be made to examine whether fewer items (i.e., the top three factor loadings) and/or a shortened response scale would be similarly informative. These data would be particularly relevant to establishing a valid and reliable integrity monitoring system that can be used by school-based personnel implementing PRAISE in the future. Indeed, the goal of community-based research is to develop measurement and intervention tools that are feasible, acceptable, sustainable, and informative in real-world settings.

Overall, findings from this study suggest that manual-based interventions can be implemented with varying levels of quality, even by the same therapists. As such, it seems that evaluating intervention integrity as it relates to program content, quality, and participant response is critical to understanding what occurs during the intervention (Perepletchikova & Kazdin, 2005). At the same time, findings related to differential outcomes based on quality of intervention delivery suggest that this aspect of intervention integrity must be explored further. In addition, consideration must be given to the context in which the intervention takes place. Not only should interventionists make clear the distinction between critical program components and areas for flexibility to fit participant needs, but context should also be considered in evaluating therapist competence (Waltz et al., 1993). Interventions that include strong integrity may result in a better understanding of how and why programs work, as well as allowing facilitators to maximize impact through adherence to critical program components in a culturally appropriate manner.

Acknowledgments

This research was supported by a research grant from the National Institute of Mental Health/National Institutes of Health (PI: Leff).

References

- Bell AJ, Borrelli B, Resnick B, Hecht J, Minicucci DS, Ory M, et al. Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH behavior change consortium. *Health Psychology* 2004;23(5):443–451. [PubMed: 15367063]
- Bilchik S. The importance of universal school-based programs in preventing violent and aggressive behavior. *American Journal of Preventive Medicine* 2007;33(2):S101–S103. [PubMed: 17675008]
- Cairns RB, Cairns BD, Neckerman HJ, Ferguson LL, Garipey J. Growth and aggression: Childhood to early adolescence. *Developmental Psychology* 1989;25:320–330.
- Crick NR. Relational aggression: The role of intent attributions, feelings of distress, and provocation type. *Development and Psychopathology* 1995;7:313–322.
- Crick NR. The role of overt aggression, relational aggression, and prosocial behavior in the prediction of children's future social adjustment. *Child Development* 1996;67:2317–2327. [PubMed: 9022243]
- Crick NR, Grotpeter JK. Relational aggression, gender, and social psychological adjustment. *Child Development* 1995;66:710–722. [PubMed: 7789197]

- Dane AV, Schneider BH. Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review* 1998;18:23–45. [PubMed: 9455622]
- Dodge KA. Social cognition and children's aggressive behavior. *Child Development* 1980;51:162–170. [PubMed: 7363732]
- Dusenbury L, Brannigan R, Falco M, Hansen WB. A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research* 2003;18(2):237–256. [PubMed: 12729182]
- Dusenbury L, Brannigan R, Hansen WB, Walsh J, Falco M. Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research* 2005;20(3):308–313. [PubMed: 15522898]
- Eisenbraun KD. Violence in schools: Prevalence, prediction, and prevention. *Aggression and Violent Behavior* 2007;12(4):459–469.
- Field, A. *Discovering statistics using SPSS*. Sage Publications; London: 2005.
- Galen BR, Underwood MK. A developmental investigation of social aggression among children. *Developmental Psychology* 1997;33:589–600. [PubMed: 9232374]
- Good T, Grouws D. Teaching effects: A process-product study of fourth grade mathematics classrooms. *Journal of Teacher Education* 1977;28:49–54.
- Gresham FM, Gansle KA, Noell GH, Cohen S, et al. Treatment integrity of school-based behavioral intervention studies: 1980–1990. *School Psychology Review* 1993;22(2):254–272.
- Kendall PC, Chu B, Gifford A, Hayes C, Nauta M. Breathing life into a manual: Flexibility and creativity with manual-based treatments. *Cognitive and Behavioral Practice* 1998;5:177–198.
- Kinney, PR.; Gray, CD. *SPSS 14 made simple*. Psychology press; New York: 2006.
- Leff, SS.; Angelucci, J.; Goldstein, AB.; Cardaciotto, L.; Paskewich, B.; Grossman, M.; Zins, J.; Elias, M.; Maher, C. *Bullying, victimization, and peer harassment: Handbook of prevention and intervention in peer harassment, victimization, and bullying*. Haworth Press; New York: 2007. Using a participatory action research model to create a school-based intervention program for relationally aggressive girls: The Friend to Friend program; p. 199-218.
- Leff SS, Cassano M, MacEvoy JP, Costigan T. Initial validation of a knowledge-based measure of social information processing and anger management. 2008 Submitted Manuscript.
- Leff SS, Costigan TE, Power TJ. Using participatory-action research to develop a playground-based prevention program. *Journal of School Psychology* 2004;42:3–21.
- Leff SS, Crick NR, Angelucci J, Haye K, Jawad AF, Grossman M, et al. Social cognition in context: Validating a cartoon-based attributional measure for urban girls. *Child Development* 2006;77(5):1351–1358. [PubMed: 16999803]
- Leff SS, Gullan RL, Paskewich B, Abdul-Kabir S, Jawad A, Grossman M, et al. An initial evaluation of a culturally-adapted social problem solving program for urban African American girls. *Journal of Prevention and Intervention in the Community*. in press.
- Leff SS, Paskewich B, Gullan RL, MacEvoy JP, Jawad A, et al. The preventing relational aggression in schools everyday (PRAISE) program: A preliminary evaluation of acceptability and impact. 2008 Manuscript in preparation.
- Leff SS, Power TJ, Manz PH, Costigan TE, Nabors LA. School-based aggression prevention programs for young children: Current status and implications for violence prevention. *School Psychology Review* 2001;30:343–360.
- Linnan, L.; Steckler, A. Process evaluation for public health interventions and research. In: Steckler, A.; Linnan, L., editors. *Process evaluation for public health interventions and research*. Jossey-Bass; San Francisco, CA: 2002. p. 1-23.
- Loeber, R.; Lacourse, E.; Hornish, DL. Homicide, violence and developmental trajectories. In: Tremblay, RE.; Hartup, WW.; Archer, J., editors. *Developmental origins of aggression*. Guilford Press; New York: 2005.
- Moncher FJ, Prinz RJ. Treatment fidelity in outcome studies. *Clinical Psychology Review* 1991;11:247–266.

- Murray-Close D, Ostrov J, Crick NR. A short-term longitudinal study of growth of relational aggression during middle childhood: Associations with gender, friendship, intimacy, and internalizing problems. *Development and Psychopathology* 2007;19:187–203. [PubMed: 17241490]
- Nastasi BK, Varjas K, Schensul SL, Silva KT, Schensul JJ, Ratnayake P. The participatory intervention model: A framework for conceptualizing and promoting intervention acceptability. *School Psychology Quarterly* 2000;15:207–232.
- National Center for Education Statistics. Indicators of school crime and safety: 2007. 2007. Retrieved December 1, 2008, from http://nces.edu.gov/programs/crimeindicators/crimeindicators2007/ind_13.asp
- Perepletchikova F, Kazdin AE. Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice* 2005;12(4):365–383.
- Perepletchikova F, Treat TA, Kazdin AE. Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology* 2007;75(6):829–841. [PubMed: 18085901]
- Peterson CA, McConnell SR. Factors affecting the impact of social interaction skills interventions in early childhood special education. *Topics in Early Childhood Special Education* 1993;13(1):38–56.
- Power TJ, Blom-Hoffman J, Clarke AT, Riley-Tillman TC, Kelleher C, Manz PH. Reconceptualizing intervention integrity: A partnership-based framework for linking research with practice. *Psychology in the Schools* 2005;42(5):495–507.
- Talbott E, Celinska D, Simpson J, Coe MG. “Somebody else making somebody else fight”: Aggression and the social context among urban adolescent girls. *Exceptionality* 2002;10(3):203–220.
- Task Force on Community Preventive Services. Effectiveness of universal school-based programs to prevent violent and aggressive behavior. *American Journal of Preventive Medicine* 2007;33(2S):S114–S129. [PubMed: 17675013]
- Waltz J, Addis ME, Koerner K, Jacobson NS. Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology* 1993;61:620–630. [PubMed: 8370857]
- Woods S, Wolke D. Direct and relational bullying among primary school children and academic achievement. *Journal of School Psychology* 2004;42:135–155.

Table 1

Descriptive statistics for process items in integrity systems 1 and 2

Integrity item	M	Median	Mode	SD	Range
System 1 process variables*					
Encourage students ^a	1.71	2.00	2	.46	1-2
Facilitator responsive ^a	1.94	2.00	2	.25	1-2
Facilitators work well together ^a	1.62	2.00	2	.53	0-2
Teachers involvement ^a	1.25	1.00	2	.76	0-2
Students' behavior ^a	1.81	2.00	2	.53	0-2
Behavior management ^a	1.81	2.00	2	.49	0-2
System 2 process variables**					
Student enthusiasm ^b	5.58	5.00	5	1.50	3-9
Student behavior ^b	5.39	5.00	6	1.48	2-8
Facilitators working ^b	5.18	5.00	5	1.19	3-7
Facilitators encourage ^b	5.48	5.00	5	1.23	3-8
Teacher participation ^b	4.18	4.00	5	1.96	1-8
Behavior management ^b	5.33	6.00	6	1.41	1-7
Facilitator enthusiasm ^b	5.15	5.00	5	1.03	3-7
Time management ^b	5.09	5.00	5	1.38	2-8
Global/overall impression ^b	5.24	5.00	6	1.42	2-8

* corresponds to 0 = Not Implemented, 1 = partially implemented, 2 = fully implemented

** corresponds to Likert scale 1-10 values, 1 = extremely poor, 5 = average/expected, 10 = extremely outstanding

^a n = 48

^b n = 33

Table 2

Bivariate correlations for the system 1 and system 2 integrity variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. System 2 student enthusiasm ^a	1.00														
2. System 2 student behavior ^a	.77***	1.00													
3. System 2 facilitators working ^a	.57**	.64***	1.00												
4. System 2 facilitators encourage ^a	.73***	.79***	.82***	1.00											
5. System 2 teacher participation ^a	.36*	.43*	.35*	.35*	1.00										
6. System 2 behavior management ^a	.48**	.73***	.54**	.68***	.47**	1.00									
7. System 2 facilitator enthusiasm ^a	.81***	.86***	.74***	.83***	.45**	.65***	1.00								
8. System 2 time management ^a	.40*	.43*	.62***	.53**	.27	.24	.52**	1.00							
9. System 2 overall impression ^a	.76***	.84***	.79***	.83***	.46**	.73***	.83***	.68***	1.00						
10. System 1 encourage students ^b	.47**	.51**	.44*	.67***	.14	.55**	.55**	.24	.54**	1.00					
11. System 1 facilitator responsive ^b	.34	.38*	.32	.21	.03	.30	.25	.18	.36*	.02	1.00				
12. System 1 facilitator work well ^b	.36*	.50**	.78***	.66***	.53**	.69***	.60***	.38*	.60***	.33*	.31*	1.00			
13. System 1 teacher involvement ^b	.30	.37*	.25	.23	.80***	.23	.41*	.15	.34	.09	.09	.19	1.00		
14. System 1 students' behavior ^b	.37*	.62***	.40*	.53**	.18	.80***	.45**	.24	.61***	.56***	.24	.42**	.12	1.00	
15. System 1 behavior management ^b	.32	.59***	.36*	.43*	.18	.82***	.13	.13	.58***	.41**	.26	.38**	.13	.84***	1.00

* $p < .05$

** $p < .01$

*** $p < .001$

^a $n = 33$

^b $n = 48$

Table 3

Factor loadings, communalities and eigenvalues from principal axis analysis of system 2 seven global integrity items

Integrity item	Factor 1: global impression	Communality
Student enthusiasm	.83	.69
Student behavior	.92	.84
Facilitators working	.83	.69
Facilitators encourage	.92	.85
Behavior management	.78	.61
Facilitator enthusiasm	.93	.87
Global/overall impression	.94	.88
Eigenvalues	5.43	
% of variance accounted for	77.49	

Note: Global student enthusiasm = student interest and enthusiasm, global student behavior = student behavior and level of distractibility, global facilitators working = facilitators working well together, global facilitators encourage = facilitators encouraging students to participate and setting up a successful intervention context, global behavior management = appropriate use of behavioral management strategies, global facilitator enthusiasm = enthusiasm of facilitators, global overall impression = overall impression of session $n = 33$

Classroom receiving intervention	Mean pre-score	Mean post-score	Mean (SD) change (Δ) score (Post-pre)*	Mean integrity factor score (#sessions)
Knowledge	6.16	9.73	3.50 ⁱ (2.94)	
HAB relational	4.82	4.00	-1.21 ^h (3.52)	
HAB overt	3.92	4.14	-.00 ^j (2.86)	
CSB relational	1.39	1.47	.06 ^k (.36)	
CSB overt	1.77	1.65	-.16 ^l (.64)	

Note: Knowledge = knowledge of social problem solving model, HAB relational = hostile attribution bias in response to relational aggression, HAB overt = hostile attribution bias in response to physical aggression, CSB relational = teacher-reported relational aggression, CSB overt = teacher reported overt aggression.

* Change scores were calculated based on all students with both posttest and pretest data

^a n = 29

^b n = 28

^c n = 30

^d n = 15

^e n = 13

^f n = 16

^g n = 18

^h n = 19

ⁱ n = 23

^j n = 22

^k n = 20

^l n = 21