Article

Evaluation & the Health Professions
1-33

DOI: 10.1177/0163278716673688
ehp.sagepub.com

Meeting the Challenges of Longitudinal Cluster-Based Trials in Schools: Lessons From the Chicago Trial of *Positive Action*

Kendra M. Lewis[1], David L. DuBois[2], Peter Ji[3], Joseph Day[4], Naida Silverthorn[2], Niloofar Bavarian[5], Samuel Vuchinich[6], Alan Acock[6], Margaret Malloy[7], Marc Schure[8] and Brian R. Flay[6]

Abstract

We describe challenges in the 6-year longitudinal cluster randomized controlled trial (CRCT) of *Positive Action* (PA), a social–emotional and character development (SECD) program, conducted in 14 low-income,

[1]Agriculture and Natural Resources, University of California, Davis, CA, USA
[2]University of Illinois at Chicago, Chicago, IL, USA
[3]Adler School of Professional Psychology, Chicago, IL, USA
[4]Governors State University, University Park, IL, USA
[5]California State University, Long Beach, Long Beach, CA, USA
[6]Oregon State University, Corvallis, OR, USA
[7]The Research Institute at Western Oregon University, Monmouth, OR, USA
[8]Montana State University, Bozeman, MT, USA

Corresponding Author:
Kendra M. Lewis, Agriculture and Natural Resources, University of California, 2801 Second Street, Davis, CA 95618, USA.
Email: kmlewis@ucanr.edu

urban Chicago Public Schools. Challenges pertained to logistics of study planning (school recruitment, retention of schools during the trial, consent rates, assessment of student outcomes, and confidentiality), study design (randomization of a small number of schools), fidelity (implementation of PA and control condition activities), and evaluation (restricted range of outcomes, measurement invariance, statistical power, student mobility, and moderators of program effects). Strategies used to address the challenges within each of these areas are discussed. Incorporation of lessons learned from this study may help to improve future evaluations of longitudinal CRCTs, especially those that involve evaluation of school-based interventions for minority populations and urban areas.

Keywords
longitudinal design, cluster randomized trial, evaluation, mobility, school based, social–emotional and character development

For youth to be successful in school and in life, they need to acquire social skills, build character, improve mental and physical health, and avoid problem behaviors (Coalition for Evidence-Based Policy, 2002; M. T. Greenberg et al., 2003). Programs designed with these goals—including those with a focus on social–emotional learning (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Weissberg & O'Brien, 2004) and positive youth development (Catalano, Berglund, Ryan, Lonczak, & Hawkins, 2004; Elias et al., 2015; Lerner, Phelps, Forman, & Bowers, 2009; Snyder & Flay, 2012)—can not only foster social and emotional skill development but also help prevent multiple problematic health-related behaviors, including bullying (Afshar & Kenny, 2008) and substance use (M. T. Greenberg et al., 2003; Zins & Elias, 2006).

Despite these encouraging findings, it is widely recognized that school-based prevention efforts present many challenges (e.g., M. T. Greenberg et al., 2003). We discuss challenges faced during a school-based prevention cluster randomized controlled trial (CRCT) relating to logistics of the study, designing the study, program fidelity, and program evaluation, and share our strategies for addressing each of challenge. Discussions of limitations associated with social intervention research are growing in popularity in the program evaluation literature (D. Greenberg & Barnow, 2014; Jaycox et al., 2006; Ong-Dean, Hofstetter, & Strick, 2010). For example, D. Greenberg and Barnow (2014) discussed eight common flaws of randomized

controlled trials (RCT) in evaluation of social programs; some of these flaws are also discussed in the present article. Additionally, Jaycox et al. (2006) discussed challenges related to recruitment, implementation, and dissemination of results. Many of the challenges and issues discussed in previous articles focus on the planning and implementation stages of RCTs. This article adds to the existing literature by considering a more comprehensive set of challenges faced in all stages of an RCT, as well as discussing some challenges not addressed in prior papers (e.g., measurement invariance in a longitudinal trial). Our goal is to provide an informative case example rather than a broad overview and synthesis of the literature. In alignment with this goal, we provide a description of issues or challenges faced, our strategy for addressing them, and sources for more in-depth discussion. Furthermore, it has been emphasized that contextual influences are critical to consider in social intervention research and that many concerns may be specific to particular settings (Trickett & Beehler, 2013). This article adds to understanding in this area by considering challenges that pertain to trials conducted in one particular type of setting—urban, low-income schools. The importance of considering challenges associated with this context is underscored by findings suggesting more limited effectiveness for programs when implemented and evaluated in low-income, urban school settings (Farahmand, Grant, Polo, Duffy, & DuBois, 2011).

The *Positive Action* (PA) Program

PA is a comprehensive, school-wide, social–emotional and character development (SECD) program grounded in theories of self-concept (Purkey, 1970; Purkey & Novak, 1970), particularly Self-Esteem Enhancement Theory (SET; DuBois, Flay, & Fagen, 2009), and is also consistent with integrative and social–ecological theories of health behaviors such as the social learning theory (Bandura, 1977), problem behavior theory (Jessor & Jessor, 1977), and theory of triadic influence (TTI; Flay & Petraitis, 1994; Flay, Snyder, & Petraitis, 2009). In line with SET, *PA* includes a classroom-based curriculum that brings into conscious awareness the motivation for self-esteem, while also teaching the skills needed for adaptive means of feeling good about oneself (e.g., self-control). In line with the TTI, a range of ecological supports (e.g., school-wide climate development, family classes) provide social reinforcement and validation for positive behaviors in both school and nonschool settings.

The *PA* program includes PreK–12 curricula, of which the K–8 portion was tested in the present trial. The scoped and sequenced classroom

curricula consist of over 140 age-appropriate lessons per grade taught for 15–20 min, 4 days/week, for Grades K–6, and 70 lessons, taught 2–3 days/week, for Grades 7 and 8. The core curricula for all grades consist of the following six units: self-concept, PAs for body (physical health) and mind (learning), social and emotional PAs focusing on getting along with others, managing, being honest with, and continually improving oneself. In addition to the student core curricula, the program includes teacher, counselor, and family training as well as school-wide climate development activities. A community component is also included in the *PA* program; however, funding for the present study did not allow for its implementation. All teachers and staff implementing *PA* in the current trial were provided 2–3 hr of on-site training at the beginning of each year.

Several evaluations have indicated significant favorable effects of *PA* on several outcomes. Most notably, two CRCTs, including the present trial, have found favorable effects for students in schools receiving *PA* on academic achievement (Bavarian et al., 2013; Snyder et al., 2010), positive youth development indicators (Lewis, Vuchinich, et al., 2016), health behaviors (Bavarian, Lewis, Acock, et al., 2016), emotional health (Lewis, DuBois, et al., 2013), self-esteem (Silverthorn et al., 2016), social environment (Bavarian, Lewis, Silverthorn, DuBois, & Flay, 2016), fewer school disciplinary incidents (Lewis, Schure, et al., 2013; Snyder et al., 2010), and lower involvement in problematic health behaviors, including substance use, violence (Lewis, Schure, et al., 2013; Li et al., 2011), and sexual activity (Beets et al., 2009).

Brief Overview of Trial

Setting, Design, School Recruitment, and Sample

The Chicago trial of *PA* sought to extend the knowledge base on the long-term effectiveness of SECD programs (Durlak et al., 2011) for urban, minority students, as well as to examine the effectiveness of such programs during the transition to adolescence. This matched-pair CRCT (Murray, 1998) of *PA* was implemented in high-poverty, K–8 schools in Chicago, with a largely minority (87%) student population. Students in seven matched pairs of schools were followed beginning in Grade 3 (Fall 2004 and Spring 2005), and at six additional times (waves) over 6 years: beginning and end of Grade 4 (Fall 2005 and Spring 2006), end of Grade 5 (Spring 2007), beginning and end of Grade 7 (Fall 2008 and Spring 2009), and end of Grade 8 (Spring 2010). The uneven distribution of waves

between Grades 5 and 8 was due to a gap in funding. Sampling and recruitment of schools took place during spring 2004 and are described in detail elsewhere (Flay, 2012; Ji, DuBois, Flay, & Brechling, 2008). Initial funding called for only five pairs (10 schools); additional funding was secured for two more pairs (14 schools total).

Measures

Student, teacher, and parent (or primary caregiver) reports were used to assess multiple outcomes throughout the duration of the CRCT. Measures were selected by a team of researchers from the sites participating in the Social and Character Development Research Consortium trial as well as members of the Consortium. Researchers and members attempted to identify measures of high reliability and validity in previous research with diverse samples of elementary school students (Social and Character Development Research Consortium, 2010). Collectively, student report measures assessed aspects of SECD (e.g., prosocial interactions, self-control, honesty), self-esteem, emotional health (e.g., anxiety), problem behaviors (e.g., substance use, violence), and school performance. Parents and teachers responded to similar measures, which also were used to obtain information about students' contexts outside of school (i.e., home and neighborhood; parent report), students' classroom behavior, academic motivation and performance, and parental school involvement (teacher report). Information on study measures can be found in the ClinicalTrials.gov posting by Flay (2012). The trial also was designed to test for intervention effects on school-level outcomes assessed via archival records of school-level absenteeism, disciplinary referrals and suspensions, and reading and math standardized test scores. These data, reported at the end of each academic year, were obtained from the Chicago Public Schools website once they became public. For the trial, we utilized these data from all study years as well as several academic years prior to the trial in order to establish a reliable baseline.

Human Subjects Approvals

This trial was approved by institutional review boards at the University of Illinois at Chicago and Oregon State University, the Research Review Board at Chicago Public Schools, and the Public/Private Ventures Institutional Review Board for Mathematica Policy Research (MPR; the latter

because the trial was part of a group of trials for which MPR collected some of the data).

Challenges and Solutions

Table 1 provides a summary of the challenges discussed in this article and strategies used to address them in the present trial. The challenges encountered in the trial are presented in four groupings: logistics, design, fidelity, and evaluation. Strategies and suggestions for addressing each challenge are incorporated into the discussion of the challenge.

Logistics of the CRCT

Recruitment of schools and assignment to the control condition. Recruiting schools to participate in an intervention trial can present a wide range of challenges. Initially, researchers may encounter difficulties that stem from having lack of established ties with schools and thus being perceived as "outsiders" (Ji et al., 2008). To help overcome such potential barriers, our strategy was to have a yearlong planning phase during which we engaged in activities that allowed us to develop rapport with school administrators (e.g., visiting schools, holding meetings). Research staff spent approximately 20 hr per school during this phase. Recruitment and matching of schools are detailed elsewhere (Ji et al., 2008). Of the 68 schools that came to an informational meeting, 18 agreed to be a part of the study on the understanding that they would be randomly assigned to either receive *PA* or be in the wait-listed control condition. Funding allowed for the inclusion of 14 schools; the remaining 4 were kept as replacement schools in the event of school attrition. Additionally, we needed to be prepared to handle the reluctance of principals to agree to participate in the trial in the event that their schools were not assigned to condition of their choice. For this trial, the control condition was the least desired, so as a strategy we offered an annual stipend to control schools to support their participation in the trial, in recognition that data collection was a substantial and potentially disruptive effort. The research team had initiated the stipend/incentive structure, not the principals of schools negotiating participation. Control schools received US$1,000 per year for each year of the trial. We also included a provision to provide the *PA* program (including training) to all control schools at no charge at the end of the trial (Ji et al., 2008). *PA* schools received a stipend of US$5,000 per year. Treatment schools had a higher stipend to support ongoing technical support for implementation and a series of onetime

Table 1. Summary of Challenges and Strategies.

Challenge	Brief Description	Strategies
Logistics		
Recruitment of schools and assignment to the control condition	Difficulty recruiting schools to participate in intensive evaluation, especially if assigned to control condition	Developed rapport with school administrators
Retention of school	Keeping all schools assigned to control condition; keeping all schools for the length of the study	Offered program to control schools at the end of the trial as well as an annual stipend; ongoing support for teachers and staff
Parental consent and parent, teacher, and student reports	Difficulty getting consent and gathering data	Monetary incentives
Assessment of student outcomes	Student self-reports may be biased	Also gathered data from other sources (parents-, teachers-, and school-level information)
Confidentiality	Survey administrations by staff may decrease confidentially and anonymity, as well as prevent students from answering honestly	Ensured confidentially and anonymity by having data collected by research staff, not teachers; asked teachers to remain seated and not walk around the room during data collection
Design		
Randomizing a small number of schools/sites	Issues with internal validity in that differences may occur between schools at baseline	Use of a matched-pair design; matched on multiple school-level variables (e.g., gender, race, and free lunch) to ensure baseline equivalency
Statistical power	Small number of clusters (schools) reduces power to find program effects	Use of a matched-pair design, longitudinal assessments, and appropriate modeling of measures to increase power

(continued)

Table 1. (continued)

Challenge	Brief Description	Strategies
Fidelity		
Implementation of PA	Fidelity of program delivery	Received buy in of teachers, not only administrators; asked teachers to complete implementation reports at the end of each unit and the end of the school year; ongoing technical support; yearly trainings; and annual meetings with other program teachers and the program developer to share experiences
Control school activities	Treatment-like activities occurring in control schools	Offered treatment at the end of the trial; offered a benign, noninterfering treatment; and assessed level of treatment-like activities
Evaluation		
Restricted range of outcomes	Data may be skewed and have "floor" or "ceiling" effects	Use of appropriate estimators to handle non-normal distributions
Measurement invariance	Construct (measure) meaning may vary	Tested measures for invariance when it was reasonable to expect invariance
Student mobility	Student population was highly mobile	Cluster-focused intent-to-treat analysis (i.e., collecting data from all students in the study schools, regardless of when they entered the trial)
Potential program moderators	Variables such as gender, race, or other characteristics may influence program effects on outcomes	Tested for moderation

Note. PA = *Positive Action*.

incentives to fund school-level activities supporting implementation. In addition to receiving program materials, treatment schools also received 2–3 hr of annual training and incentives or tokens of appreciation for completion of assessments on students and implementation reports.

Retention of schools. The 18 schools that agreed to be in the study were randomized and no schools were lost because of randomization to the control condition (Ji et al., 2008). In the event of school attrition, four schools were retained throughout the trial. However, we did not lose any schools during the trial, thus giving us a school level $N = 14$. This accomplishment seems likely to be attributable, in part, to the above-described strategies and solutions (e.g., rapport, stipends). We also had a research coordinator (generally a graduate student) to support schools in the study. All schools signed a memorandum of understanding (MOU); school personnel that were new to the schools during the trial would also sign the MOU. Finally, we also emphasized maintaining positive working relationships with schools, particularly administrators and key support staff, and recommend the use of this strategy. This included periodic check-ins, problem-solving flexibly to accommodate the needs and preferences of individual schools with respect to activities such as structuring of program implementation and scheduling of data collection sessions and providing thank-you cards and small tokens of appreciation (e.g., food treats for teachers and other staff) at key junctures throughout the trial. As with the planning phase, approximately 20 hours per school was spent on these activities.

Consent and survey completion rates. Gaining consent from a high proportion of parents from urban areas or of ethnic minority is often a challenge, potentially because of a lack of parent involvement (National Center for Educational Statistics, 2004), risk level of the sample (Rojas, Sherrit, Harris, & Knight, 2008), or cultural differences between the researchers and participants (Rodríguez, Rodríguez, & Davis, 2006). In the present trial, we obtained parental consent at the start of the study when students were in Grade 3 (assent from students was also required at each wave, but did not prove to be a significant challenge for this trial, as assent rates averaged more than 95%). At baseline, parents of 79% of students provided consent. Parental consent for students joining the study at later waves was obtained at those times; these consent rates ranged from 65% to 78% for Waves 2–5. It was also necessary to reconsent all students at Wave 6 based on receipt of additional funding which allowed the original trial to be extended. Consent

rates were lower at this latter stage of the study (≈58–64% for Waves 6–8), which is consistent with previous research indicating a drop in rate of consent at higher grade levels (Ji, Pokorny, & Jason, 2004; Thompson, 1984). Parental consent that was obtained lasted through the length of the study. That is, parents did not have to provide consent at every wave. One key limiting factor in achieving a high rate of consent in school-based intervention research is simply getting forms returned from parents. In order to address this challenge, our strategy was to provide incentives. Parents were offered US$10 for returning consent forms in early waves, regardless of their consent decisions. While consent rates were higher likely as a result of this incentive, we made a point of communicating that the incentives were not conditional on a "yes" decision, but rather simply returning a completed form. We also offered pizza parties for classrooms (and gift certificates for teachers) with 90% or higher returned consent forms (Ji et al., 2006). Other strategies include establishing a relationship with school personnel, having student assistants to help contact parents, and having clear consent forms (Blom-Hoffman et al., 2009; Fletcher & Hunter, 2003).

To help ensure high rates of survey completion among consented students and their parents and teachers, we offered financial incentives for completion of their respective surveys at each wave (US$20 for parents, which was increased to US$40 at the final wave; US$50 for teachers; and a US$5 coupon to a local restaurant for students, which was increased to US$10 at later waves). The increased incentives for parents increased survey completion rates (e.g., 50% at Wave 5 for a US$20 incentive to 73% at Wave 8 for a US$40 incentive).

Assessment of student outcomes. A majority of the student-level outcomes were assessed via self-report, potentially leading to a method bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Self-reports in particular are susceptible to social desirability biases; students might exaggerate their participation in high-risk behaviors in order to feel as if they fit in with their peers. Alternatively, they might underreport such behaviors knowing society's negative views on behaviors such as substance use and bullying. Of additional concern is the possibility that students in *PA* schools might report more desirable behaviors because these are emphasized as a key goal of the program; such a tendency could artificially inflate estimates of program effects. Other research suggests that such biases are of minimal concern (Bachman, Johnston, & O'Malley, 1996; Elliott, 1994; Krohn, Thornberry, Gibson, & Baldwin, 2010; Spoth et al., 2007). As one strategy for addressing this possibility, care was taken during data collection

throughout the trial to provide instructions that encouraged students to respond honestly and that normalized reporting of undesirable behaviors. Specifically, we had the following instructions at the beginning of the survey: ''You give the answer that is MOST TRUE FOR YOU. Please remember that there are no right or wrong answers—we just want your honest opinions.'' As a further strategy to guard against possible self-report bias, we supplemented with teacher and parent reports. These reports had strong internal consistency; however, the correlations with student reports on similar outcomes were low to moderate. There are a few possible reasons for this. Teachers do not know their students well at the beginning of the school year, so their ratings at that time are subject to lack of both reliability and validity for this reason, thus compromising the utility of any change scores based on them. Also, students have different teachers each year and, in addition, for school in the current trial, teacher turnover was also high. As for parents, given their relatively low levels of formal education, the potential also exists for their reports to have less than optimal levels of reliability and validity. This is particularly so given that strategies most likely to be useful in countering such a possibility (e.g., personal interviews) were unable to be incorporated into our trial due to resource constraints. Finally, teachers and parents are rating behaviors based on what they see and where they see it (e.g., school and home, respectively), which are different environments and also differ from the student's experience. So while these reports may be reliable for short-term changes, they are not as reliable for looking at behavior change as in this trial. We recommend supplementing with the previously noted use of school-level archival records data as additional measures of intervention impact.

A related issue concerns the confidentiality of student survey responses. To the extent that students do not feel safe in the privacy of their responses, for example, they may fail to share sensitive information (e.g., involvement in substance use or other risk behaviors). To address this concern, research staff, rather than teachers, administered surveys, and assurance of confidentially of responses were part of the assent process. Additionally, we used a system of coding ID numbers on surveys rather than having names. To further ensure confidentially for students, teachers were asked to stay seated at their desks and not walk around the classroom during survey administration. This also reduced teacher burden in that they were not responsible for data collection. Future researchers may also want to consider electronic data collection such as utilizing school computer labs or iPads as these may be useful in addressing confidentiality concerns of students (e.g., needing to

hand in a paper survey that includes their responses to sensitive questions), while also easing the burden of survey collection.

Design of the CRCT

Randomizing a small number of schools/sites. The small number of clusters (i.e., schools) in this study presents concerns both for statistical conclusion validity (i.e., low statistical power; Shadish, Cook, & Campbell, 2002) and the potential that randomizing a small number of units would leave differences between treatment and control schools at baseline (Murray, Varnell, & Blitstein, 2004). Other threats included school-level attrition and failed randomization. To aid in randomization, we used a matched-pair design, a form of blocking in that the blocks are matched pairs. Exact matching can be difficult; but the use of a distance-matrix method for matching (Schochet & Novak, 2003) led to very tight matching in this case. Lack of equivalence on covariates at baseline could decrease power and precision, thereby creating a threat to statistical conclusion validity. Matching and establishing baseline equivalency reduces variability within pairs. Similar to the randomized block design, this approach reduces variability within treatment conditions and potential confounding, producing a better estimate of treatment effects (Imai, King, & Nall, 2009; Ivers et al., 2012; Rhodes, 2014).

Matching was implemented using multiple school-level archival variables including ethnicity, attendance rate, truancy rate, number of students per grade, information about school crime rates, percentage of parents reported to demonstrate school involvement, percentage of teachers employed by the school who met minimal teaching standards and percentage of students who met or exceeded criteria for passing the Illinois State Achievement Test (ISAT), received free lunch, and enrolled in or left school during the academic year (Ji et al., 2008). The use of a range of measures in the matching process helped to ensure baseline equivalency between treatment and control groups on school- and student-level outcomes. Equivalency tests at the school level revealed no statistically significant differences between the treatment and control group schools on any of the matching variables at baseline or at any of the three other times tested (Ji et al., 2008; Lewis, Bavarian, et al., 2012).

Furthermore, although random assignment from matched pairs of clusters (e.g., schools) cannot guarantee equivalence of the nested subjects (e.g., students; Giraudeau & Rivaud, 2009), of the total 74 student-, parent-, and teacher-reported scales tested for baseline equivalency, only 8 showed significant differences, 4 favoring control students and 4 favoring *PA* students.

Table 2. Minimal Detectable Effect Sizes in a Cluster Randomized Controlled Trial.

ICC	Cluster ($N = 10$)		Cluster ($N = 14$)	
	Two-tailed p	One-tailed p	Two-tailed p	One-tailed p
.01	.35	.29	.28	.24
.05	.52	.43	.42	.35
.10	.67	.56	.54	.46
.15	.79	.66	.64	.54
.20	.90	.75	.72	.61

Note. ICC = intraclass correlation. All estimates are derived using the Optimal Design software (Spybrook et al., 2011) assuming .80 as a desired level of statistical power and a Type I error rate of .05. Cluster sized of 10 was the original cluster size based on funding, additional funding allowed for four more clusters for a final cluster size of 14.

The low number of statistically significant differences and their varying directions suggest that the matching and randomization were successful and that threats to internal validity were minimized.

Statistical power. Due to funding agency decisions and funding constraints, only 14 schools could be included in the trial (increased from an initial 10 schools required by the funder). Current literature suggests that a sample size of seven school pairs would usually not be adequate for multilevel modeling (Hox, 2010) or for detecting small or moderate effects (Bingenheimer & Raudenbush, 2004). Our strategies of a matched-pairs design, analytic procedures appropriate to the distributions of outcomes, and assessment of outcomes over repeated occasions all served to help improve the level of statistical power (Raudenbush, Martinez, & Spybrook, 2007; Shadish et al., 2002) in addition to securing more funding to increase the number of schools from 10 to 14. Table 2 displays the minimum detectable effect size (MDES; Bloom, 1995) for the difference between intervention and control schools in multilevel models under several conditions. These estimates were calculated using the Optimal Design (Plus Version 3.0) software (Spybrook et al., 2011) and the "Cluster Randomized Trials with person-level outcomes" and "Repeated measures" options within the program. As shown in Table 2, MDES values are smaller for the larger N, smaller intraclass correlations (ICCs), and one- versus two-tailed tests. Increasing the number of clusters (i.e., schools) from 10 to 14 improved the two-tailed MDES values to levels similar to those for one-tailed values with $N = 10$. Improving power to be adequate for two-tailed tests of significance was particularly desirable given concerns about using

one-tailed tests in assessing intervention effects (Ringwalt, Paschall, Gorman, Derzon, & Kinlaw, 2011). A number of methodological factors (e.g., number of students per group, repeated measures correlations) are important considerations in planning CRCTs (for a detailed discussion, see Murray et al., 2004).

Fidelity

Implementation of PA. Getting teachers to deliver a prevention program curriculum regularly and with fidelity also can present an array of formidable challenges (e.g., organizational capacitiy, staff characteristics; Durlak & DuPre, 2008; Malloy et al., 2014). Further, many factors play a role in program implementation, such as integration into school activities, training, and supervision of the program (Payne & Eckert, 2009). Given the increasing demands on schools and teachers to focus on traditional academic subject instruction in particular, we anticipated that the support of school administrators and social support among colleagues would likely both be critical to ensuring implementation of the *PA* program.

With this concern in mind, our strategy was to ask principals to obtain the buy in of teachers before agreeing to participate in the study. Providing more specific expectations for principals' efforts in this regard and asking for documentation of some requisite level of teacher buy in would have been desirable, and is recommended for future trials. We provided an incentive (e.g., a free lunch) to teachers to regularly complete brief implementation reports (completed at the end of each unit of the *PA* program, about every 6 weeks). A member of the research team worked with the program developer to provide ongoing technical support for implementation (e.g., visits to schools approximately every 2 weeks). Additionally, representatives from all schools were brought together annually to share experiences with each other (e.g., professional learning communities; Mullen & Schunk, 2010) and the program developer. Finally, as noted earlier, the program developer provided yearly staff trainings for each school. Training for the program was a challenge as well. We took advantage of in-service days and also paid for substitutes using grant funds. The amount of training was based on what the schools would accommodate and therefore likely compromised fidelity of implementation to some unknown degree. Resources for implementation support limited our capacity to ensure that 100% of teachers and other staff participated in trainings or, if absent or later arriving to the school, received make-up training. However, by structuring the training as part of required in-service/preparation

days, attendance was very high. Furthermore, when a teacher or staff person joined a school midyear, the implementation coordinator made every effort to meet individually with them to orient them to the program and cover content similar to that covered in the beginning-of-year trainings. The periodic trainings conducted during each school year provided further opportunities for training.

Both teacher reports and student reports were used for the process evaluation. Teachers were asked about the number of lessons they taught and how much they made adaptations to the curriculum, if at all. Students rated their engagement with the program ("I like *PA*"; "I plan to use PAs when I grow up"). We attempted to obtain weekly reports of implementation activities (e.g., *PA* lessons taught) from teachers in treatment schools, but this was found to be too burdensome, so we relied on the six unit reports (e.g., self-concept, PAs for body, etc.). In addition to teacher reports, we recommend using observations or taping sessions to assess implementation quality. We did not use these methods in the present trial, as this was logistically too difficult to utilize effectively within the resource constraints of the study and the contexts of the participating schools.

Consistent with the issues and dynamics we had anticipated, an implementation study found that teachers who had delivered *PA* reported that although they saw SECD programs as beneficial, they found it difficult to prioritize implementing the *PA* program given pressures emanating from regional administrators and the school system's central office to devote classroom instructional time to academics. As we had expected, teachers also reported that support from colleagues was critical to implementation (Fagen et al., 2015).

More generally, indices pertaining to implementation (e.g., teacher description of amount and quality of *PA* activities in the classroom, perceived effectiveness of the activities, student reports of exposure to and attitudes toward the program) tended to show variability across schools, especially in early years, with improvements over time (Bickman et al., 2009). By the end of Year 6, one school was implementing at only a moderate level of fidelity, three at a moderate to high level, and three at high levels (Jarpe-Ratner et al., 2013). Although neither the levels nor the consistency of implementation achieved are ideal, it seems likely that shortcomings in this aspect of the trial would have been notably more pronounced in the absence of the above-described strategies.

Control school activities. The transition from laboratory to field settings is not a smooth one (Hulleman & Cordray, 2009), and there are a variety of

sources of infidelity that can impact the theoretically expected differences between treatment and control conditions. One such source of infidelity often ignored is what happens in control schools (e.g., Sloboda et al., 2008). A requirement of study participation was that schools had not previously utilized *PA* or a similar SEL/SECD intervention, so that the *PA* program effects would not be confounded with other programs. However, neither this provision nor offering the *PA* program to control schools upon completion of the study period served to prevent some control schools from using some *SECD-oriented* activities similar to those of the *PA* program during the study period; such a prohibition would, in fact, raise serious ethical considerations and likely would have inhibited school recruitment efforts.

Teachers completed surveys about whether they used SECD-like activities in their classroom, including specifics on the target domain (e.g., peace promotion, character education), program name (if they used a program), strategies at the classroom and school level for promoting SECD, and attitudes toward promoting SECD. These surveys were completed annually; the completion rate was over 85% at all waves (Social and Character Development Research Consortium, 2010). Some control schools did indeed report using SECD-oriented activities. We do not know the vigor of the activities (level of implementation, school support, etc.), other than that control schools reported a high quantity of SECD-oriented activities. However, treatment teachers reported engaging in more SECD-like activities and professional development than control teachers, as well as reported higher enthusiasm for these activities than did control teachers (Social and Character Development Research Consortium, 2010). It is possible, then, that the estimated effects of the *PA* program, as indexed by differences on measures between conditions (i.e., effect sizes), that have been reported are understated because of the SECD-related activities occurring in control schools.

We recommend assessing the level of treatment-like activities in control schools and including this stipulation in an MOU with schools. By doing so, we were able to have some idea of the extent of this bias. In principal, such assessments could then be utilized in testing models of intervention impact, such as complier average causal effects models (Stuart, Perry, Le, & Ialongo, 2008), although the present trial lacked a sufficient number of schools to do so.

Evaluation

Restricted range of outcomes. Most questionnaires were from existing, validated measures. Some had to be modified, informed by pilot testing and a

general concern with item content being appropriate for students in a range of grades (from third to eighth). Therefore, some censoring or "floor" and "ceiling" effects (restricted range) were observed for some of these measures (e.g., respect for parents). In addition, other outcomes were relatively rare (e.g., extreme forms of violence). As a result, for some measures the standard assumptions for a basic linear regression type analysis did not hold. Using such an approach can lead to misleading estimates of program effects (Long & Freese, 2006). Our strategy and solution for this issue has been to utilize generalized linear mixed models appropriate to the distribution of each outcome. Researchers should be prepared to analyze data using methods appropriate to the distribution beyond a normal distribution, such as binary or categorical (Cohen, Cohen, West, & Aiken, 2013; Long & Freese, 2006; Rabe-Hesketh & Skrondal, 2012), Poisson (Olsen & Schafer, 2001; Rabe-Hesketh & Skrondal, 2012), and censored (Joreskog, 2002; Meng & Schenker, 1999; Skrondal & Rabe-Hesketh, 2004) models.

Measurement invariance. Measurement invariance establishes that the construct (measure) meaning is similar for groups (e.g., boys and girls, conditions) or across time (e.g., age; Geiser, 2012; Geldhof & Stawski, 2016; Pentz & Chou, 1994). Many measures, however, are designed to be stable over time. Measurement invariance may not be a reasonable expectation in a longitudinal study when a developmental change is expected as well as change due to intervention effects for the treatment group (Geldhof et al., 2014). In this trial, our analyses to date have revealed several measures to not exhibit strong invariance over times of measurement (Lewis, Vuchinich, et al., 2016). We have no a priori reason, however, to expect that such noninvariance was differentially applicable to the intervention condition, such as might occur if *PA* affected youths' interpretation of items, a consideration which supports the meaningfulness of our treatment effect estimates (Geldhof et al., 2014). Additionally, given the comprehensive nature of *PA*, measures were used that provided relatively global assessments of constructs (e.g., overall tendencies toward negative affect rather than assessments of particular types of feelings). Such measures seem less subject to problematic variation over time than those that are more specific and thus more prone to developmental change. If measurement invariance does not hold, groups cannot be directly compared, as the items have different meaning between the groups; these qualitative distinctions in the items should be explored and discussed (e.g., the intervention may have changed the treatment group's understanding of a construct). An in-depth discussion of measurement invariance is beyond the scope of this article; however, we

recommend that researchers test measures for invariance when it was reasonable to expect invariance. Further, researchers are recommended to test for changes due to development versus program effects (Geldhof & Stawski, 2016; Lawrence & Blair, 2003; Pentz & Chou, 1994).

Student mobility. Although all 14 schools were retained throughout the trial, the student population attending these schools was highly mobile. Within low-income, urban schools, mobility is a common occurrence (Tobler & Komro, 2011). Student mobility poses challenges for efficacy trials of longitudinal school-based programs. One such challenge includes difficulty inferring whether the observed effects are due to the intervention or a result of differential attrition. A strategy proposed to overcome these disadvantages is to conduct a cluster-focused intent-to-treat analysis (Brown et al., 2008; Vuchinich, Flay, Aber, & Bickman, 2012). This approach acknowledges the focus on schools and follows all schools randomized to condition to trial endpoint, regardless of how well the intervention is implemented (or not) in treatment schools. It also requires collecting and analyzing data from all students who are in the appropriate grade cohort in the schools at the time of each assessment. In accordance with these considerations, we assessed students who entered schools after the beginning of the trial (joiners), but did not follow individual students who stopped attending the study schools (leavers). This is similar to the repeated cross-sectional design used in community-level research (Murray, 1998), except that we surveyed the population of students present at each time rather than taking a sample of them.

From the standpoint of students, across time they could be considered a "dynamic" (i.e., changing) grade cohort. Table 3 shows the number of students present at each wave, as well as the number of students who remained in later waves of the trial; there were no differences by condition. At the student level, the total sample size across all eight waves was 1,170 with approximately half that number being present at each wave. Only 21% (131) of the original 624 Grade 3 cohort remained at Grade 8, illustrating the high mobility of the low-income, urban students in our sample. Additionally, as noted above, parental consent rates declined during the middle-school grades; furthermore, simultaneously, Chicago school enrollment was decreasing during the time of the study. For all of these reasons, the student sample size at Wave 8 (363) was smaller than at Wave 1 (624). The average number of waves of data provided per student was 3.1.

Table 3. Mobility of Students by Study Condition.

Wave	1	2	3	4	5	6	7	8
Season/Year	Fall/ 2004	Spring/ 2005	Fall/ 2005	Spring/ 2006	Spring/ 2007	Fall/ 2008	Spring/ 2009	Spring/ 2010
Grade	3	3	4	4	5	7	7	8
Treatment								
Number of participating students	316	306	244	297	262	104	187	195
Number of joiners		39	0	77	47	0	83	50
Number of leavers		49	62	24	82	158	0	42
Control								
Number of participating students	308	299	220	244	253	92	172	168
Number of joiners		27	0	55	60	0	81	27
Number of leavers		36	79	31	51	161	1	31
Totals	624	605	464	541	515	196	359	363
Number of joiners		66	0	132	107	0	164	77
Number of leavers		85	141	55	133	319	1	73

Note. The increase in mobility rates after Wave 5 may be partially explained by the time difference between Wave 5 and Wave 6 representing one school year plus two summer breaks of mobility (end of Grade 5 to beginning of Grade 7) and the transition from elementary to middle school grades. Joiners in the Fall of 2005 and 2008 were considered as joiners at the end of the school year (Spring 2006 and 2009, respectively). This table shows the analysis sample. There may be small variations between this table and tables in papers (and reports) from Institute of Education Sciences and the Social and Character Development Research Consortium because some students may have only been present for the "multisite" or "site-specific" days of data collection. Additionally, some students may not have been present for any data collection and therefore only have teacher-reported data.

Potential moderators of program effectiveness. Programs like *PA* may have different effects for boys and girls, for students at different levels of risk, or for students as well as entire schools that differ in other ways. Moderation analyses are necessary to test for such possibilities, and these are now

becoming more common in the literature (e.g., Bierman, Nix, Greenberg, Blair, & Domitrovich, 2008). Indeed, several programs have found gender differences in social–emotional outcomes and health-related behaviors (e.g., Bierman et al., 2010; Flay, Graumlich, Segawa, Burns, & Holliday, 2004; Taylor, Liang, Tracy, Williams, & Seigle, 2002). Other potential moderators include implementation (e.g., Durlak et al., 2011; Wilson & Lipsey, 2007), behavioral risk (e.g., Ellickson, McCaffrey, Ghosh-Dastidar, & Longshore, 2003; Wilson & Lipsey, 2007), and environmental factors (e.g., Bierman et al., 2010; Hughes, Cavell, Meehan, Zhang, & Collie, 2005). In addition, race or ethnicity moderation may be of interest. It is important to consider potential moderators prior to data collection to ensure that these moderators are assessed in the measurement tools.

To date, in the Chicago trial of *PA*, gender differences have been found on only a limited and inconsistent basis (Bavarian et al., 2013; Lewis, Schure, et al., 2013). Further, using student race or ethnicity as a moderator was not possible because of extreme confounding with aggregate racial/ethnic composition at the school level. Median odds ratio (MOR) for Black versus non-Black schools ranged from 6.39 to 20.30 (ICCs = .53 to .75), showing that student race/ethnicity largely covaried with school (MOR is preferable to ICCs with binary outcomes; Merlo et al., 2006). Indeed, three pairs of schools were >99% Black, two pairs were 75% Hispanic, and two pairs were mixed (50% Hispanic, 31% Black, 9% White, and 9% Asian). Ideally, we would test for moderation by factors such as race/ethnic student composition, baseline levels of student academic achievement and attendance, and so forth at the school level. However, with only 14 total schools in the trial, such analyses are not feasible due to their extremely low statistical power. To date, no moderation analyses by implementation or risk factors have been assessed in the present trial.

Of particular interest in highly mobile populations, program effects might have differed based on whether students stayed in the same school for the duration of the program, left the school during the study, or joined the school during the study. Because the trial involved random assignment at the cluster (school) level, data were collected not only from students present at the start of the trial but also those who joined the school after the trial began. We thus have been able to explore whether program impacts differed depending on student mobility pattern, as indicated by whether data were present or missing for a student at each wave. A challenge, however, is that there are 256 possible patterns of student mobility (present or absent at each wave to the eighth power = 256).

An approach to analyzing missing data (mobility) patterns is latent class analysis (LCA; Beunckens, Molenberghs, Verbeke, & Mallinckrodt, 2008; Lin, McCulloch, & Rosenheck, 2004; Roy, 2003), as LCA allows for the identification of classes of students with similar patterns of missingness (Marsh, Lüdtke, Trautwein, & Morin, 2009). Drawing on this work, we conducted an LCA to identify subgroups of individuals (Flay, 2012; Lewis, Bavarian, et al., 2017). The results of this analysis revealed five distinct patterns of student mobility during the trial: (1) stayers (average study duration of 5.72 years, 13%), (2) temporary participants (present for Grade 4 and/or 5 only; average study duration of 1.30 years; 16%), (3) late joiners (average study duration of 1.38 years; 25%), (4) early leavers (average study duration of .94 years; 22%), and (5) late leavers (average study duration of 3.23 years; 24%). We have used these patterns as a grouping variable to test for mobility as a moderator of program effects—that is, examining whether program effects varied by mobility pattern.

To date, these analyses have not revealed significant differences in estimated program effects by mobility pattern (e.g., Lewis, Vuchinich, et al., 2016). Ideally, researchers also should collect and utilize data on auxiliary variables as these aid in the imputation process (Collins, Schafer, & Kam, 2001; Enders, 2008; Graham, 2003) that may help to predict or explain patterns of missingness or mobility in school-based CRCTs. For this trial, information was collected on several family variables that could help to explain mobility, including prior moving history, job stability, employment status, and the likelihood of moving during the next year. In future studies, we plan to incorporate these variables into our analyses. Additionally, presence or absence at a particular wave is based on whether the student had any data on any outcome at the wave. The student may have been absent on the days of data collection; that is, they may have technically been "present" that school year or wave, but just not for data collection. A more accurate approach would be to use attendance data; however, only aggregated (not individual student) data was not available for our use. Researchers and evaluators in schools may want to examine the possibility of gaining access to student-level records such as attendance for various analyses.

Discussion

Although a number of challenges were encountered in our school-based cluster randomized, we were able to address each of them (at least partially) using practical strategies that may be useful in other school-based CRCTs. It is worth noting that with these strategies incorporated, the trial was able to

meet many of the standards put forth by the Society of Prevention Research for standards of evidence for prevention programs and policies (Flay et al., 2005; Gottfredson et al., 2015) such as efficacy Standard 2 regarding measurement of the outcomes and follow-up, and efficacy Standard 5 regarding the reporting of all outcomes, regardless of direction or significance. Further, this program is included in the What Works Clearinghouse, Blueprints for Healthy Youth Development, and CrimeSolutions.gov as a model program. In line with these strengths, multiple papers reporting the intervention impact from this trial have been accepted for publication in a wide range of peer-reviewed journals (Bavarian, Lewis, Acock, et al., 2016; Bavarian et al., 2013; Lewis, Bavarian, et al., 2012; Lewis, DuBois, et al., 2013; Lewis, Schure, et al., 2013; Lewis, Vuchinich, et al., 2016). Our experience suggests that moving to an effectiveness trial in low-income schools for a multifaceted, and thus relatively complex, intervention such as *PA* would be likely to present serious and intensified challenges to achieving an adequate level of fidelity. To help offset this concern, it could be desirable to build into the intervention itself some of the external supports that were incorporated into the current trial (e.g., implementation coordinator to work with the schools that is affiliated with the program, not the school).

In this trial, we were quite successful in addressing a number of challenges that were unique to, or at least present in more pronounced ways, within urban, low-income schools. School-based CRCTs of prevention and promotion programs such as *PA* can be expected to play a critical role as we move to effectiveness and dissemination trials using more population- or setting-based approaches (Flay, 1986; Glasgow, Lichtenstein, & Marcus, 2003). Longitudinal CRCTs face many challenges, and the goal of this article was to elucidate several of these challenges and corresponding strategies for addressing them. The strategies described were not fully successful in all instances. Illustratively, the levels of implementation fidelity achieved were clearly not ideal. Likewise, not all potential strategies discussed could be utilized in this trial due to inherent limitations such as the relatively small number of schools involved. Lastly, to reiterate, this article does not provide a broad synthesis of the literature of CRCTs and the challenges faced in such trials. Such a paper would be informative, however, and clearly would benefit from greater availability of papers such as the present one that provide in-depth consideration of issues faced in specific trials. In this way, it will be possible to build a robust knowledge base to inform not only the science but also the "art" of conducting rigorous and informative CRCTs in different settings.

Authors' Note

Kendra M. Lewis, Niloofar Bavarian, Marc Schure, and Margaret Malloy were affiliated with Oregon State University during initial preparation of this article. Joseph Day was with University of Illinois at Chicago. The Social and Character Development (SACD) research program includes multiprogram evaluation data collected by MPR and complementary research study data collected by each grantee. The findings reported here are based only on the Chicago portion of the multiprogram data and the complementary research data collected by the University of Illinois at Chicago and Oregon State University (Brian Flay, Principal Investigator) under the SACD program. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Institute of Education Sciences, CDC, MPR, or every Consortium member, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Correspondence concerning this article should be addressed to Kendra Lewis or Brian Flay.

Acknowledgments

We would like to thank Robert Duncan for assistance with analyses. We are extremely grateful to the participating Chicago Public Schools (CPS), their principals, teachers, students, and parents. We thank the CPS Research Review Board and Office of Specialized Services, especially Drs. Renee Grant-Mitchell and Inez Drummond, for their invaluable support of this research.

Declaration of Conflicting Interests

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research described herein was conducted using the program, the training, and technical support of *Positive Action*, Inc. in which Brian Flay's spouse holds a significant financial interest. Issues regarding conflict of interest were reported to the relevant institutions and appropriately managed following the institutional guidelines.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was funded by grants from the Institute of Education Sciences (IES), US Department of Education: R305L030072, R305L030004, and R305A080253 to the University of Illinois, Chicago (2003-05) and Oregon State University (2005-12). The initial phase (R305L030072), a component of the Social and Character Development (SACD) Research Consortium, was a collaboration among IES, the Centers for Disease Control and Prevention's (CDC) Division of Violence Prevention, Mathematica Policy Research Inc. (MPR), and awardees of SACD cooperative agreements (Children's Institute, New York University, Oregon State University, University at

Buffalo-SUNY, University of Maryland, University of North Carolina-Chapel Hill, and Vanderbilt University).

References

Afshar, P., & Kenny, M. (2008). Violence in schools, cross-national and cross-cultural perspectives. *Journal of Psychological Trauma*, *6*, 87–89. doi:10.1080/19322880802096624

Bachman, J. G., Johnston, L. D., & O'Malley, P. M. (1996). *The Monitoring the Future project after twenty-two years: Design and procedures* (Monitoring the Future occasional paper no. 38, p. 57). Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bavarian, N., Lewis, K. M., Acock, A., DuBois, D. L., Yan, Z., Vuchinich, S., … Flay, B. R. (2016). Effects of a school-based social–emotional and character development program on health behaviors: A matched-pair, cluster-randomized controlled trial. *The Journal of Primary Prevention*, *37*, 87–105. doi:10.1007/s10935-016-0417-8

Bavarian, N., Lewis, K. M., DuBois, D. L., Acock, A., Vuchinich, S., Silverthorn, N., … Flay, B. R. (2013). Using social-emotional and character development to improve academic outcomes: A matched-pair, cluster-randomized controlled trial in low-income, urban schools. *Journal of School Health*, *83*, 771–779. doi:10.1111/josh.12093

Bavarian, N., Lewis, K. M., Silverthorn, N., DuBois, D. L., & Flay, B. R. (2016). *Impact of a school-based social-emotional program on social environments: Results from a trial in low-income schools*. Paper presented at the Society for Prevention Research 24th Annual Meeting, San Francisco, CA.

Beets, M. W., Flay, B. R., Vuchinich, S., Snyder, F. J., Acock, A., Li, K.-K., … Durlak, J. (2009). Use of a social and character development program to prevent substance use, violent behaviors, and sexual activity among elementary-school students in Hawaii. *American Journal of Public Health*, *99*, 1438–1445. doi:10.2105/ajph.2008.142919

Beunckens, C., Molenberghs, G., Verbeke, G., & Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal gaussian data. *Biometrics*, *64*, 96–105. doi:10.1111/j.1541-0420.2007.00837.x

Bickman, L., Riemer, L. M., Brown, J. L., Jones, S. M., Flay, B. R., Li, K.-K., … Massetti, G. (2009). Approaches to measuring implementaion fidelity in school-based program evaluations. *Journal of Research in Character Education*, *7*, 75–102.

Bierman, K. L., Coie, J. D., Dodge, K. A., Greenberg, M. T., Lochman, J. E., McMahon, R. J., & Pinderhughes, E. (2010). The effects of a multiyear universal

social–emotional learning program: The role of student and school characteristics. *Journal of Consulting and Clinical Psychology*, *78*, 156–168. Retrieved from http://dx.doi.org/10.1037/a0018607

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start Redi program. *Development and Psychopathology*, *20*, 821–843. doi:10.1017/S0954579408000394

Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: Issues in the application of multilevel models. *Annual Review of Public Health*, *25*, 53–77. doi:10.1146/annurev.publhealth.25.050503.153925

Blom-Hoffman, J., Leff, S. S., Franko, D. L., Weinstein, E., Beakley, K., & Power, T. J. (2009). Consent procedures and participation rates in school-based intervention and prevention research: Using a multi-component, partnership-based approach to recruit participants. *School Mental Health*, *1*, 3–15. doi:10.1007/s12310-008-9000-7

Bloom, H. S. (1995). Minimum detectable effects. *Evaluation Review*, *19*, 547–556. doi:10.1177/0193841X9501900504

Brown, C. H., Wang, W., Kellam, S. G., Muthen, B. O., Petras, H., Toyinbo, P., ... Windham, A. (2008). Models for testing and evaluating impact in randomized field trials: Intent-to-treat analyses for integrating the perspectives of person, place, and time. *Drug and Alcohol Dependence*, *95S*, S74–S104. doi:10.1016/j.drugalcdep.2007.11.013

Catalano, R. F., Berglund, M. L., Ryan, J. A. M., Lonczak, H. S., & Hawkins, J. D. (2004). Positive youth development in the united states: Research findings on evaluations of positive youth development programs. *The Annals of the American Academy of Political and Social Science*, *591*, 98–124. doi:10.1177/0002716203260102

Coalition for Evidence-Based Policy. (2002). *Bringing evidence-driven progress to education: A recommended strategy for the U.S. Department of Education*. Washington, DC: J. Baron.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York, NY: Routledge.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351. doi:10.1037//1082-989X.6.4.330

DuBois, D. L., Flay, B. R., & Fagen, M. C. (2009). Self-esteem enhancement theory: An emerging framework for promoting health across the life-span. In R. J. DiClemente, M. C. Kegler, & R. A. Crosby (Eds.), *Emerging theories in health promotion practice and research* (2nd ed., pp. 97–130). San Francisco, CA: Jossey-Bass.

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*, 327–350. doi:10.1007/s10464-008-9165-0

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*, 405–432. doi:10.1111/j.1467-8624.2010.01564.x

Elias, M. J., Leverett, L., Duffell, J., Humphrey, N., Stepney, C. T., & Ferrito, J. J. (2015). Integrating social-emotional learning with related prevention and youth-development approaches. In C. D. J. A. Durlak, R. P. Weissberg, & T. P. Gullota (Ed.), *Handbook of social and emotional learning: Research and practice* (pp. 39–49) New York, NY: Guilford.

Ellickson, P. L., McCaffrey, D. F., Ghosh-Dastidar, B., & Longshore, D. L. (2003). New inroads in preventing adolescent drug use: Results from a large-scale trial of project alert in middle schools. *American Journal of Public Health*, *93*, 1830–1836. doi:10.2105/AJPH.93.11.1830

Elliott, D. S. (1994). Serious violent offenders: Onset, developmental course, and termination—The American Society of Criminology 1993 Presidential Address. *Criminology*, *32*, 1–21. doi:10.1111/j.1745-9125.1994.tb01144.x

Enders, C. K. (2008). A note on the use of missing auxilliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling*, *15*, 434–448. doi:10.1080/10705510802154307

Fagen, M. C., Gilmet, K., McRae, K., Patten, V., DuBois, D. L., Allred, C. G., & Flay, B. R. (2015). *Exploring teacher perspectives on implementation of a school-based social-emotional and character development program*. Manuscript in preparation.

Farahmand, F. K., Grant, K. E., Polo, A. J., Duffy, S. N., & DuBois, D. L. (2011). School-based mental health and behavioral programs for low-income, urban youth: A systematic and meta-analytic review. *Clinical Psychology: Science and Practice*, *18*, 372–390. doi:10.1111/j.1468-2850.2011.01265.x

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, *15*, 451–474. doi:10.1016/0091-7435(86)90024-1

Flay, B. R. (2012). *Randomized trial of the Positive Action program in chicago schools and extension to grade 8*. Retrieved from http://clinicaltrials.gov/show/NCT01025674

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., … Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175. doi:10.1007/s11121-005-5553-y

Flay, B. R., Graumlich, S., Segawa, E., Burns, J. L., & Holliday, M. Y. (2004). Effects of 2 prevention programs on high-risk behaviors among African American youth: A randomized trial. *Archives of Pediatrics & Adolescent Medicine*, *158*, 377–384. doi:10.1001/archpedi.158.4.377

Flay, B. R., & Petraitis, J. (1994). The theory of triadic influence: A new theory of health behavior with implications for preventive interventions. *Advances in Medical Sociology*, *4*, 19–44. Retrieved from http://www.emeraldinsight.com/books.htm?issn=1057-6290

Flay, B. R., Snyder, F., & Petraitis, J. (2009). The theory of triadic influence. In R. J. DiClemente, M. C. Kegler, & R. A. Crosby (Eds.), *Emerging theories in health promotion practice and research* (2nd ed., pp. 451–510). San Francisco, CA: Jossey-Bass.

Fletcher, A. C., & Hunter, A. G. (2003). Strategies for obtaining parental consent to participate in research. *Family Relations*, *52*, 216–221. doi:10.1111/j.1741-3729.2003.00216.x

Geiser, C. (2012). *Data analysis with Mplus*. New York, NY: Guilford Press.

Geldhof, G. J., Bowers, E. P., Johnson, S. K., Hershberg, R., Hilliard, L. J., & Lerner, R. M. (2014). Relational developmental systems theories of positive youth development: Methodological issues and implications. In P. C. Molenaar, R. M. Lerner, & K. Newell (Eds.), *Handbook of developmental systems theory and methodology* (pp. 66–94). New York, NY: Guilford.

Geldhof, G. J., & Stawski, R. S. (2016). Invariance. In S. K. Whitbourne (Ed.), *The encyclopedia of adulthood and aging* (pp. 689–694). Hoboken, NJ: Wiley.

Giraudeau, B., & Rivaud, P. (2009). Preventing bias in cluster randomized trials. *PLoS Medicine*, *6*, e1000065. doi:10.1371/journal.pmed.1000065

Glasgow, R. E., Lichtenstein, E., & Marcus, A. C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health*, *93*, 1261–1267. doi:10.2105/AJPH.93.8.1261

Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*, 893–926. doi:10.1007/s11121-015-0555-x

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling*, *10*, 80–100. doi:10.1207/S15328007SEM1001_4

Greenberg, D., & Barnow, B. S. (2014). Flaws in evaluations of social programs illustrations from randomized controlled trials. *Evaluation Review*, *5*, 359–387. doi:10.1177/0193841X14545782

Greenberg, M. T., Weissberg, R. P., O' Brien, M., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning *American Psychologist*, *58*, 466–474. doi:10.1037/0003-066X.58.6-7.466

Hox, J. J. (2010). *Multilevel analysis*. London, England: Routledge.

Hughes, J. N., Cavell, T. A., Meehan, B. T., Zhang, D., & Collie, C. (2005). Adverse school context moderates the outcomes of selective interventions for aggressive children. *Journal of Consulting and Clinical Psychology*, *73*, 731–736. Retrieved from http://dx.doi.org/10.1037/0022-006X.73.4.731

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*, 88–110. doi:10.1080/19345740802539325

Imai, K., King, G., & Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, *24*, 29–53. doi:10.1214/08-STS274

Ivers, N. M., Halperin, I. J., Barnsley, J., Grimshaw, J. M., Shah, B. R., Tu, K., . . . Zwarenstein, M. (2012). Allocation techniques for balance at baseline in cluster randomized trials: A methodological review. *Trials*, *13*, 120–128. doi:10.1186/1745-6215-13-120

Jarpe-Ratner, E., Fagen, M., Day, J., Gilmet, K., Prudowsky, J., Neiger, B. L., . . . Flay, B. R. (2013). Using the community readiness model as an approach to formative evaluation. *Health Promotion Practice*, *5*, 649–655. doi:10.1177/1524839913487538

Jaycox, L. H., McCaffrey, D. F., Ocampo, B. W., Shelley, G. A., Blake, S. M., Peterson, D. J., . . . Kub, J. E. (2006). Challenges in the evaluation and implementation of school-based prevention and intervention programs on sensitive topics. *American Journal of Evaluation*, *27*, 320–336. doi:10.1177/1098214006291010

Jessor, R., & Jessor, S. L. (1977). *Problem behavior and psychosocial development*. New York, NY: Academic Press.

Ji, P., DuBois, D. L., Flay, B. R., & Brechling, V. (2008). "Congratulations, you have been randomized into the control group!(?)": Issues to consider when recruiting schools for matched-pair randomized control trials of prevention programs. *Journal of School Health*, *78*, 131–139. doi:10.1111/j.1746-1561.2007.00275.x

Ji, P., Flay, B. R., DuBois, D. L., Brechling, V., Day, J., & Cantillon, D. (2006). Consent form return rates for third-grade urban elementary students. *American Journal of Health Behavior*, *30*, 467–474. Retrieved from http://dx.doi.org/10.5993/AJHB.30.5.3

Ji, P., Pokorny, S. B., & Jason, L. A. (2004). Factors influencing middle and high schools' active parental consent return rates. *Evaluation Review*, *28*, 578–591. doi:10.1177/0193841X04263917

Joreskog, K. (2002). *Censored variables and censored regression*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/censor.pdf

Krohn, M. D., Thornberry, T. P., Gibson, C. L., & Baldwin, J. M. (2010). The development and impact of self-report measures of crime and delinquency. *Journal of Quantitative Criminology*, *26*, 509–525. doi:10.1007/s10940-010-9119-1

Lawrence, F. R., & Blair, C. (2003). Factorial invariance in preventive intervention: Modeling the development of intelligence in low birth weight, preterm infants. *Prevention Science*, *4*, 249–261. doi:10.1023/A:1026068115471

Lerner, J. V., Phelps, E., Forman, Y., & Bowers, E. P. (2009). Positive youth development. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology* (Vol. 1, 3rd ed., pp. 524–558). Hoboken, NJ: John Wiley and Sons.

Lewis, K. M., Bavarian, N., Duncan, R., Acock, A., DuBois, D. L., Silverthorn, N., … Flay, B. R. (2017). *Student mobility patterns in a longitudinal, school-based, cluster-randomized cohort prevention trial*. Manuscript in preparation.

Lewis, K. M., Bavarian, N., Snyder, F. J., Acock, A., Day, J., DuBois, D. L., … Flay, B. R. (2012). Direct and mediated effects of a social-emotional and character development program on adolescent substance use. *International Journal of Emotional Education*, *4*, 56–78. Retrieved from http://www.enseceurope.org/journal/

Lewis, K. M., DuBois, D. L., Bavarian, N., Acock, A., Silverthorn, N., Day, J., … Flay, B. (2013). Effects of positive action on the emotional health of urban youth: A cluster-randomized trial. *Journal of Adolescent Health*, *53*, 706–711. doi:10.1016/j.jadohealth.2013.06.012

Lewis, K. M., Schure, M. B., Bavarian, N., DuBois, D. L., Day, J., Ji, P., … Flay, B. R. (2013). Problem behavior and urban, low income youth: A randomized controlled trial of *Positive Action* in Chicago. *American Journal Of Preventive Medicine*, *44*, 622–630. Retrieved from http://dx.doi.org/10.1016/j.amepre.2013.01.030

Lewis, K. M., Vuchinich, S., Ji, P., DuBois, D. L., Acock, A., Bavarian, N., … Flay, B. R. (2016). Effects of the *Positive Action* program on indicators of positive youth development among urban youth. *Applied Developmental Science*, *20*, 16–28. doi:10.1080/10888691.2015.1039123

Li, K.-K., Washburn, I., DuBois, D. L., Vuchinich, S., Ji, P., Brechling, V., … Flay, B. R. (2011). Effects of the *Positive Action* programme on problem behaviors in elementary school students: A matched-pair, randomized control trial in chicago. *Psychology & Health*, *26*, 187–204. doi:10.1080/08870446.2011.531574

Lin, H., McCulloch, C. E., & Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics*, *60*, 295–305. doi:10.1111/j.0006-341X.2004.00173.x

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. College Station, TX: Stata Press.

Malloy, M., Acock, A., DuBois, D. L., Vuchinich, S., Silverthorn, N., Ji, P., & Flay, B. R. (2014). Teachers' perceptions of school organizational climate as predictors of dosage and quality of implementation of a social-emotional and character development program. *Prevention Science*, *16*, 1086–1095. doi:10.1007/s11121-014-0534-7

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. S. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person- and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling*, *16*, 191–225. doi:10.1080/10705510902751010

Meng, X., & Schenker, N. (1999). Maximum likelihood estimation for linear regression models with right censored outcomes and missing predictors. *Computational Statistics & Data Analysis*, *29*, 471–483. Retrieved from http://dx.doi.org/10.1016/S0167-9473(98)00074-7

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., … Larsen, K. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Health*, *60*, 290–297. doi:10.1136/jech.2004.029454

Mullen, C. A., & Schunk, D. H. (2010). A view of professional learning communities through three frames: Leadership, organization, and culture. *McGill Journal of Education/Revue des sciences de l'éducation de McGill*, *45*, 185–203. doi:10.7202/045603ar

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, *94*, 423–432. doi:10.2105/AJPH.94.3.423

National Center for Educational Statistics. (2004). *The condition of education*. Washington, DC: U.S. Department of Education.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, *96*, 730–745. doi:10.1198/016214501753168389

Ong-Dean, C., Hofstetter, C. H., & Strick, B. R. (2010). Challenges and dilemmas in implementing random assignment in educational research. *American Journal of Evaluation*, *32*, 29–49. doi:10.1177/1098214010376532

Payne, A. A., & Eckert, R. (2009). The relative importance of provider, program, school, and community predictors of the implementation quality of school-based prevention programs. *Prevention Science*, *11*, 126–141. doi:10.1007/s11121-009-0157-6

Pentz, M. A., & Chou, C.-P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology*, *62*, 450. Retrieved from http://dx.doi.org/10.1037/0022-006X.62.3.450

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903. doi:10.1037/0021-9010.88.5.879

Purkey, W. W. (1970). *Self-concept and school achievement*. Englewood Cliffs, NJ: Prentice-Hall.

Purkey, W. W., & Novak, J. (1970). *Inviting school success: A self-concept approach to teaching and learning*. Belmont, CA: Wadsworth.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata* (3rd ed., Vol. II). College Station, TX: Stata Press.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*, 5–29. doi:10.3102/0162373707299460

Rhodes, W. (2014). Pairwise cluster randomization: An exposition. *Evaluation Review*, *38*, 217–250. doi:10.1177/0193841X14540654

Ringwalt, C., Paschall, M. J., Gorman, D., Derzon, J., & Kinlaw, A. (2011). The use of one-versus two-tailed tests to evaluate prevention programs. *Evaluation & the Health Professions*, *34*, 135–150. doi:10.1177/0163278710388178

Rodríguez, M. D., Rodríguez, J., & Davis, M. (2006). Recruitment of first-generation latinos in a rural community: The essential nature of personal contact. *Family Process*, *45*, 87–100. doi:10.1111/j.1545-5300.2006.00082.x

Rojas, N. L., Sherrit, L., Harris, S., & Knight, J. R. (2008). The role of parental consent in adolescent substance use research. *Journal of Adolescent Health*, *42*, 192–197. Retrieved from http://dx.doi.org/10.1016/j.jadohealth.2007.07.011

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, *59*, 829–836. doi:10.1111/j.0006-341X.2003.00097.x

Schochet, P., & Novak, T. (2003). *Computer program to construct school pairs*. Unpublished manuscript.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.

Silverthorn, N., DuBois, D. L., Lewis, K. M., Reed, A., Bavarian, N., Day, J., … Flay, B. R. (2016). *Effects of a school-based social-emotional and character development program on self-esteem levels and processes: A cluster-randomized controlled trial*. Manuscript under review.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. London, England: Chapman & Hall/CRC Press.

Sloboda, Z., Pyakuryal, A., Stephens, P. C., Teasdale, B., Forrest, D., Stephens, R. C., & Grey, S. F. (2008). Reports of substance abuse prevention programming available in schools. *Prevention Science*, *9*, 276–287. doi:10.1007/s11121-008-0102-0

Snyder, F. J., & Flay, B. R. (2012). Positive youth development. In A. Higgins-D'Alessandro, M. Corrigan, & P. Brown (Eds.), *The Handbook of Prosocial Education* (pp. 415–443). Lanhan, MD: Rowman and Littlefield.

Snyder, F. J., Flay, B. R., Vuchinich, S., Acock, A., Washburn, I., Beets, M. W., & Li, K. K. (2010). Impact of a social-emotional and character development program on school-level indicators of academic achievement, absenteeism, and disciplinary outcomes: A matched-pair, cluster-randomized, controlled trial. *Journal of Research on Educational Effectiveness*, *3*, 26–55. doi:10.1080/19345740903353436

Social and Character Development Research Consortium. (2010). *Efficacy of school-wide programs to promote social and character development and reduce problem behavior in elementary school children*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Spoth, R., Redmond, C., Shin, C., Greenberg, M., Clair, S., & Feinberg, M. (2007). Substance-use outcomes at 18 months past baseline: The prosper community–university partnership trial. *American Journal of Preventive Medicine*, *32*, 395–402. doi:10.1016/j.amepre.2007.01.014

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal design plus empirical evidence: Documentation for the "Optimal Design" software. Retrieved from http://wtgrantfoundation.org/library/uploads/2015/11/OD-Documentation-V3.pdf

Stuart, E. A., Perry, D. F., Le, H.-N., & Ialongo, N. S. (2008). Estimating intervention effects of prevention programs: Accounting for noncompliance. *Prevention Science*, *9*, 288–298. doi:10.1007/s11121-008-0104-y

Taylor, C. A., Liang, B., Tracy, A. J., Williams, L. M., & Seigle, P. (2002). Gender differences in middle school adjustment, physical fighting, and social skills: Evaluation of a social competency program. *Journal of Primary Prevention*, *23*, 259–272. doi:10.1023/A:1019976617776

Thompson, T. L. (1984). A comparison of methods of increasing parental consent rates in social research. *Public Opinion Quarterly*, *48*, 779–787. doi:10.1086/268883

Tobler, A. L., & Komro, K. A. (2011). Contemporary options for longitudinal follow-up: Lessons learned from a cohort of urban adolescents. *Evaluation and Program Planning*, *34*, 87–96. doi:10.1016/j.evalprogplan.2010.12.002

Trickett, E. J., & Beehler, S. (2013). The ecology of multilevel interventions to reduce social inequalities in health. *American Behavioral Scientist*, *57*, 1227–1246. doi:10.1177/0002764213487342

Vuchinich, S., Flay, B., Aber, L., & Bickman, L. (2012). Person mobility in the design and analysis of cluster-randomized cohort prevention trials. *Prevention Science*, 1–14. doi:10.1007/s11121-011-0265-y

Weissberg, R. P., & O'Brien, M. U. (2004). What works in school-based social and emotional learning programs for positive youth development. *The Annals of the American Academy of Political and Social Science*, *591*, 86–97. doi:10.1177/0002716203260093

Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine*, *33*, S130–S143. doi:10.1016/j.amepre.2007.04.011

Zins, J. E., & Elias, M. J. (2006). Social and emotional learning. In G. G. Bear & K. M. Minke (Eds.), *Children's needs III: Development, prevention, and intervention* (pp. 1–13). Bethesda, MD: National Association of School Psychologists.