MEASURING BEHAVIORAL OUTCOMES ASSOCIATED WITH COMPREHENSIVE CHARACTER EDUCATION PROGRAMS
A Practical Approach to Using Fewer Schools in School-Randomized Controlled Trials While Maintaining Adequate Statistical Power

Carol K. Holtzapple
Department of Research, The Flippen Group

Character education programs support the development of positive character traits in children and adults. Effective violence prevention programs improve pro-social competencies and reduce negative behaviors in students by enhancing protective factors (strong bonds with teachers; clear rules of conduct that are consistently enforced) and targeting modifiable risk factors (inappropriate behavior; poor social coping skills). Comprehensive character education or prevention programs implemented in schools are those programs that affect a school as a community; thus, they are considered to be school-level interventions. Randomization of schools is the preferred method of choice when evaluating school-wide interventions; however, studies using this design are very costly because of the large number of schools generally required to provide the statistical power (precision) needed to obtain meaningful results. The purpose of this report is to demonstrate how prestudy archival school data combined with a stratification approach to analysis can be used to reduce the number of schools that are required in studies involving comprehensive, school-wide interventions. Obtaining baseline information from school archival data for stratification purposes prior to initiating a study assists researchers, school districts, and other stakeholders in choosing study sites (prior to randomization) that will provide the greatest likelihood that changes in behavioral outcomes can be measured if they occur.

• **Carol K. Holtzapple,** Director of Research, The Flippen Group, 1199 Haywood Dr., College Station, TX 77845. E-mail: carol.holtzapple@flippengroup.com

Journal of Research in Character Education, 9(1), 2011, pp. 57–69 ISSN 1543-1223

INTRODUCTION

Character education programs address moral and ethical values such as respect, responsibility, trustworthiness, and caring concern for others (Battistich, 2003; Berkowitz & Beir, 2006a, 2006b; Flay, Allred, & Ordway, 2001; Greenberg, Kusché, Cook, & Quamma, 1995; Lickona & Davidson, 2005; What Works Clearinghouse, 2007). Positive character development is also a focus of related program areas such as social emotional learning (Elias & Arnold, 2006; Zins, Elias, Greenberg, & Weissberg, 2000), drug abuse prevention (Battistich, Schaps, Watson, Solomon, & Lewis, 2000; NIDA, 2002) and violence prevention (Aber, Brown, & Jones, 2003; Office of Juvenile Justice and Delinquency Prevention, 2004).

Although some programs may be designed to address positive character development in individual students, comprehensive programs address the school environment that impacts students' social and emotional health as well as their academic success. When an intervention is likely to produce effects in all study participants (intervention and control) within a school, studies that attempt to randomize at the teacher level will tend to underestimate the intervention's effect. For instance, high school students change teachers and classrooms throughout the day. When school administrators are able to provide training for only half of the teachers within a school, it is difficult to measure behavioral changes in students who spend part of the day with both trained (intervention) and untrained (control) teachers. In addition, intervention teachers may communicate to control teachers the skills they learn in the training; thus, students and teachers can create a situation in which "spillover" of knowledge occurs between intervention and control groups. Therefore, for studies of comprehensive (school-wide) character education interventions in which a spillover effect is likely to be large, randomization of schools is the preferred research design (Bloom, Richburg-Hayes, & Black, 2007).

Unfortunately, studies using school-randomized designs are often very costly because of the large number of schools typically required to provide the statistical precision or power needed to obtain meaningful results (Bloom et al., 2007). Program outcomes are reported in terms of effect sizes (experimental—control differences divided by the standard deviation) so that research results across studies can be compared (Best Evidence Encyclopedia, 2011). The effect size, δ, as well as the number of schools (clusters), J, impact the magnitude of the statistical power (Raudenbush & Liu, 2000; Raudenbush, Spybrook, Liu, & Congdon, 2004; Spybrook, Raudenbush, Liu, & Congdon, 2009) of school-randomized trials. An increase in either parameter will increase statistical power. However, the number of schools (J) that can be used in a district-wide study is often constrained by budgetary or logistical issues. Therefore, the expected effect size can play a determining role in the decision to commit district resources for a randomized controlled trial to investigate the effectiveness of character education programs in local schools.

The magnitude of the effect size (δ) is dependent upon two distinct parameters (Spybrook et al., 2006, 2009): (1) the amount of change that occurs as a result of the intervention being investigated (the difference between the experimental and control outcomes), and (2) the variability in the prestudy data (standard deviation). Although the former parameter is strongly determined by the effectiveness of the intervention, the latter is determined by the data already recorded in school archival data obtained in the years prior to the study. This data can be incorporated into the study design such that researchers have greater control over the level of prestudy variability that exerts a strong impact on the magnitude of the effect size (Holtzapple, 2009; Raudenbush Martinez, & Spybrook, 2007).

Minimizing prestudy variability is an integral part of study design in other fields of research (i.e., chemical analysis) as researchers use laboratory instruments whose variances

are strictly controlled by the manufacturer. As a result, much larger effect sizes than those typically observed in educational research studies can be obtained using very few samples. For instance, in the field of drug-residue analysis, an effect size of 2.0 standard deviation (*SD*) units (*ES* = 2.0) is generally required by journals when reporting the lowest limits of detection for chemical contaminants (Holtzapple, Buckley, & Stanker, 1997; Holtzapple, Pishko, & Stanker, 2000). In these types of research studies, low variability and tight control of the experimental design is possible. For research studies in the social sciences, Cohen (1988) recognized that effect sizes are likely to be smaller due to difficulties inherent in measuring outcomes in studies involving high variability and lower control over experimental designs. He suggested effect size labels of "small" (*ES* ~ 0.2), "medium" (*ES* ~ 0.5), and "large" (*ES* ~ 0.8) when reporting outcomes in the social sciences.

What Works Clearinghouse states that reports of cluster-level (i.e., school-level) effect sizes are relatively rare, that there is not yet enough knowledge in the field to judge the magnitude of cluster-level effect sizes, and that Cohen's effect size categories that are appropriate for reporting outcomes when individual-level assignment is used can be misleading when applied to studies employing cluster-level assignment (Valentine & Cooper, 2003; What Works Clearinghouse [WWC], 2008). This would suggest that researchers who develop cluster-randomized trials involving comprehensive program models cannot expect to use Cohen's designations of small (~ 0.2) to large (~ 0.8) effect sizes when working with cluster-randomized trials.

In a recent review of 55 cluster-randomized studies funded by the National Center for Education Research (NCER; Spybrook, 2008), recomputed minimum detectable effect size (MDES) values ranged from approximately 0.10 to more than 1.70. Three studies with much higher recomputed MDES values were excluded from further analysis. It is possible that large MDES values obtained from single studies or from reviews of multiple studies are routinely excluded when reporting results in the literature because they are unusual in social science research. However, it is of great value to determine whether or not these higher MDES values are appropriate to use when working with cluster-randomized trials. School districts are required to document the effects of comprehensive character education or prevention programs in local schools, but districts typically cannot manage more than 6 to 10 schools in cluster-randomized studies. Therefore, in educational trials using few schools, larger MDES values (similar to those required in chemical trials) must be attained in order for studies to have adequate power.

In the present report, power analyses were used to determine the feasibility of developing randomized, controlled trials using fewer than 10 schools (clusters). The variability in the prestudy, 3-year data involving the number of discipline referrals—an important, school-level outcome that is investigated when implementing effective comprehensive character education (WWC, 2008) and prevention (Center for Disease Control, 2010) programs—was addressed prior to randomization in order to meet the statistical demands imposed on the study due to the limited number of schools (clusters) used in the research trial. Parallels between the effect sizes required for adequate statistical power in drug residue detection studies and in cluster-randomized educational trials involving fewer than 10 clusters are discussed.

METHODS

General Research Design

School administrators from the Oneida-Herkimer-Madison Board of Cooperative Educational Services (BOCES) in New York and the Riverside County Office of Education (RCOE) in California were committed to measuring the effect of a comprehensive, school-based program designed to improve

prosocial competencies and reduce negative behaviors in students. In order to meet federal and state guidelines for demonstrating program effectiveness in local schools using evidence-driven research, an experimental (randomized, controlled), blocked, research design was used, with schools chosen as the unit of randomization. Public high schools served by BOCES and RCOE were asked to provide data from the 3 years preceding the study (SY2005, 2006, & 2007). After stratification of the schools on the basis of coefficient of variation values derived from the 3-year discipline referral data, eight schools were selected for the study, paired according to school type (grades 7-12 or grades 9-12), and then randomly assigned from the matched pairs to intervention or control groups.

Data Collection

Schools were provided a prestudy data form requesting demographic, academic, and behavioral data for the 3 school years preceding initiation of the study. Principals from each of the schools submitted the data to their respective leadership teams. For this particular program that impacts school climate through strengthening teacher-student connectedness, the BOCES and RCOE leadership teams were primarily interested in measuring differences in behavioral outcomes (i.e. reduction of risk factors such as discipline referrals and acquisition of protective factors such as pro-social relational skills).

Statistical Analysis

Statistical power analyses were conducted using the Optimal Design software (initially version 1.76 and then the updated version 2.0) for cluster randomized trials (Spybrook et al., 2006; Spybrook et al., 2009). The significance level (α) was held constant at 0.05 for all calculations. Univariate ANOVA analyses were conducted using SPSS Version 16.0 software (SPSS, 2010), with the independent variable being the application of the intervention in the schools and the dependent variable being the number of discipline referrals per year.

RESULTS

Power Analysis

School districts are often limited in the number of schools that are available to participate in randomized, controlled trials; in many cases, decision makers may not have the information they need to determine whether or not studies will be able to measure changes in school outcomes. Power analyses can provide this information by indicating what the minimum effect size needs to be for any given number of schools such that meaningful results can be obtained.

The power analyses provided in Figure 1 demonstrate how large the effect sizes need to be in order for studies involving 6, 8, 10, 20 or 40 schools to have adequate power (0.8) when considering a randomized, controlled, blocked, random effects model. In the Optimal Design software (v. 2.0; available from the W.T. Grant Foundation, 2011), the following power analysis design components were chosen: (a) cluster randomized trial with cluster level outcome (measurement of group processes); (b) multi-site (or blocked) cluster randomized trial; (c) treatment at level 2; (d) power for the treatment effect on the *y*-axis; and (e) power versus effect size (delta). The following parameters within the designated power analyses design were then used to generate "best case scenario" (low variability/high reliability) power curves: number of clusters (J) = 6, 8, 10, 20, or 40; effect size variability (σ^2) = 0.01 (random effects model); number of sites (K) = J/2; reliability (rel) = 0.9; and proportion of explained variance by the blocking variable (B) = 0.15. The x-axis ($\leq x \leq$) was expanded in order to view minimum detectable effect size (MDES) values from 0 to 2.5.

As can be seen in Figure 1, when 40 or 20 schools are considered, the MDES values (at power = 0.8) for school-level randomized

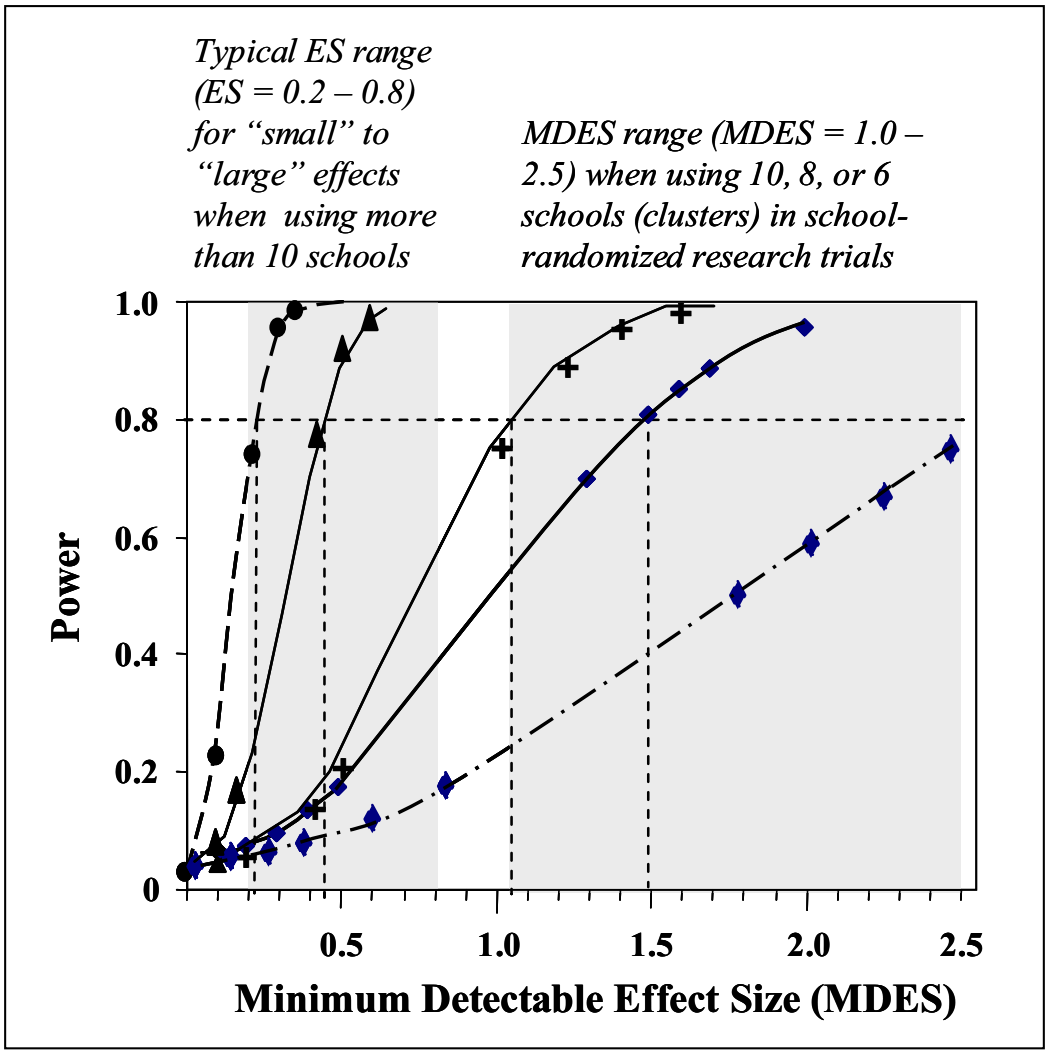

FIGURE 1

Power analyses to determine minimum effect sizes (MDES) required for adequate statistical power. Parameters for all curves were set as follows: α = 0.05, $K = J/2$ (sites), reliability (rel) = 0.9; and proportion of explained variance by the blocking variable (B) = 0.15. Computed minimum detectable effect size (MDES) values when $J = 40$ (-●-), 20 (-▲-), 10 (-+-), 8 (- -) or 6 (-♦-) are 0.22, 0.45, 1.07, 1.49, and 2.80, respectively.

studies fall within the MDES range associated with individual-level assignment (approximately 0.2 for small effect sizes to 0.8 for large effect sizes; Cohen, 1988). However, when 10, 8, or 6 schools are considered, Figure 1 provides clear supportive evidence for the assertion by WWC (2008) that reporting effect size values normally associated with individual-level assignment can be misleading. Even in the best-case scenario (low variability and high cluster-level reliability) presented in Figure 1, the minimum detectable effect size (MDES) required to attain sufficient power (0.8) is greater than 1.0 when using 10 or fewer school clusters. Although effect sizes of this magnitude would be considered to be very

large when using individual-level assignment, they are merely adequate for achieving sufficient statistical power when reporting study results using cluster-level assignment, cluster-level outcomes, and 10 or fewer clusters.

This requirement of larger MDES values when working with fewer than 10 schools parallels the requirement in drug residue detection studies that the difference between the mean outcome of the control and the mean outcome of the "lowest detectable (residue) limit" must be equal to 2 standard deviation units of the control (MDES = 2.0). Typically, in drug residue detection studies, 3 to 4 replicates are used for the control and standard samples in order to develop standard curves for drug residue detection. This is analogous to using 3 to 4 schools per treatment group (and 3 to 4 schools per control group), which corresponds to J (number of schools) = 6 to 8. Figure 1 demonstrates that the MDES value (at power = 0.8) for J = 6 is greater than 2.5 whereas the MDES value for J = 8 is approximately 1.5 when high cluster-level reliability and low effect size variability parameters are considered. Because of the much higher MDES value associated with using only 6 clusters, $J = 8$ was chosen as a test case to determine whether or not a theoretical MDES value of ~ 1.5 can be reasonably attained in whole-school research studies in practice.

The data in Figure 1 suggests that school districts would be able to measure outcomes even if only 8 schools can participate in a study. However, the data assumes a best case scenario, which rarely occurs in practice. This being the case, power analyses were undertaken to determine the range of expected minimum detectable ES values that would be required in studies in which the variability and reliability within and between schools are suboptimal. The power analyses shown in Figure 2 demonstrate how the effect size variability, cluster-level reliability, and proportion of the explained variance by the blocking variable affect the MDES value when $J = 8$ (4 control and 4 intervention schools) and $K = 4$ (4 pairs of matched schools). The following parameters were used to generate the power curves for $A_{8\ (\text{best case scenario})}$, B_8 and C_8: number of clusters (J) = 8; number of sites (K) = J/2; effect size variability (σ^2) = 0.01 (A_8), 0.1 (B_8), or 0.2 (C_8) using the random effects model; reliability (rel) = 0.9 (A_8), 0.7 (B_8), or 0.5 (C_8); and the proportion of the explained variance by the blocking variable (B) = 0.15 (A_8), 0.10 (B_8), or 0 (C_8). The x-axis (≤x≤) was expanded in order to view MDES (δ) values from 0 to 2.5.

The results demonstrate that the minimum detectable ES values range from a best case scenario of 1.49 *(A_8: $J = 8$, $K = 4$, $\sigma^2 = 0.01$, reliability = 0.9, $B = 0.15$)* to a suboptimal scenario of 2.33 *(C_8: $J = 8$, $K = 4$, $\sigma^2 = 0.2$, reliability = 0.5, $B = 0$)* when parameters that impact the magnitude of the effect size are varied while keeping the number of schools (J = 8) and sites (K = 4) constant. An MDES value of 1.49 to 2.33 (or -1.49 to -2.33 when a decrease in negative behavioral outcomes is desired) indicates that for a given outcome in a school-randomized study, the mean outcome observed in the intervention schools must differ from that observed in the control schools by at least 1.49 standard deviation (SD) units.

The results involving MDES values obtained using the Optimal Design statistical software demonstrate that, for instances in which financial or logistical constraints limit the number of schools that can be used in research, districts can theoretically conduct worthwhile studies of comprehensive (school-wide) character education programs. However, researchers must abandon the statistical parameters typically used in the social sciences (large number of samples, high variability, and comparatively small acceptable ES values) and use parameters that are typical of the physical sciences (small number of samples, low variability, and relatively high ES values). Either set of parameters is valid, but researchers cannot mix the two systems. If fewer than 10 schools are used (i.e., a physical sciences parameter), then as What Works Clearinghouse (2008) suggests, it would be misleading to report that $ES = 0.5$ (i.e., a social sciences parameter) represents a "medium"

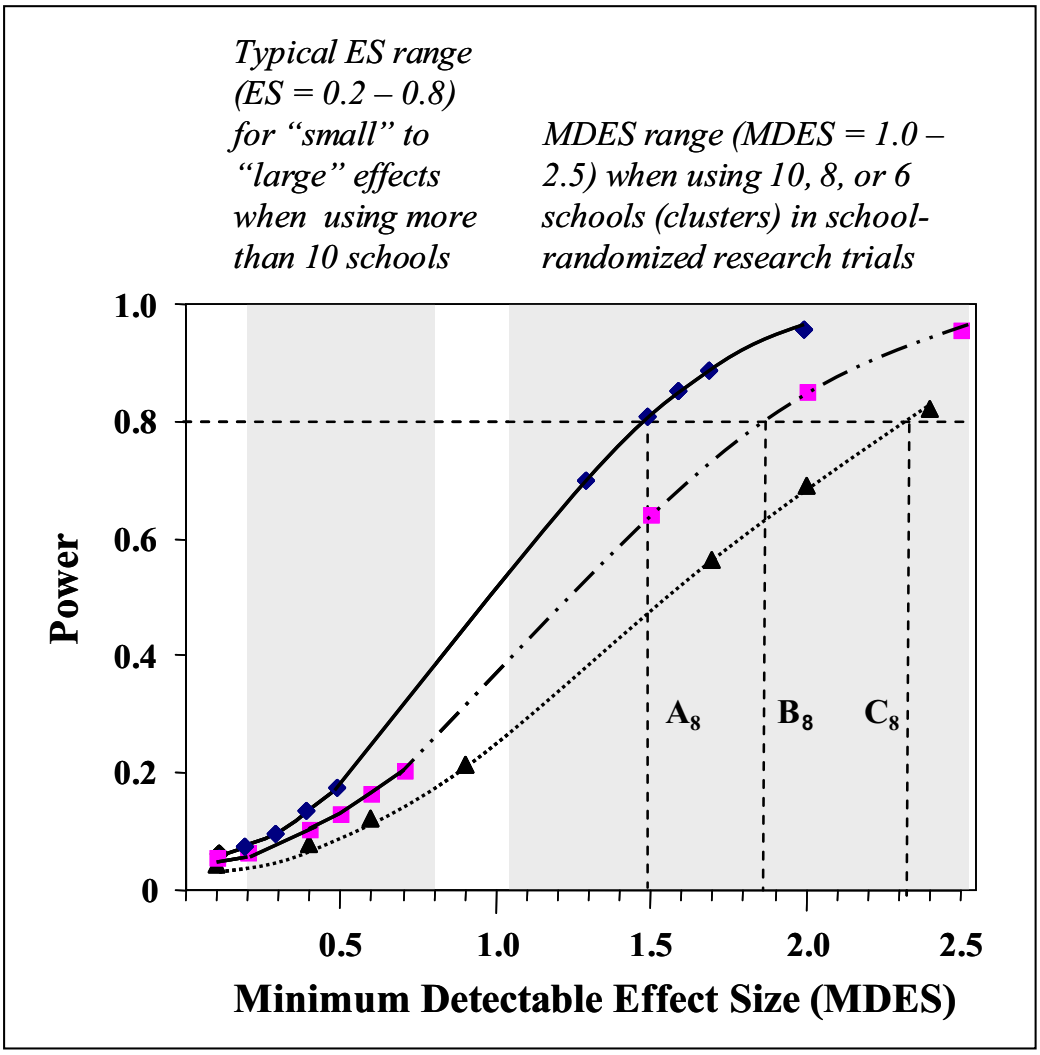

FIGURE 2

Power analyses to determine minimum effect sizes required for adequate statistical power for an 8-school study while varying other parameters. The following parameters were used to generate power curves A_8 (- -), B_8 (-■-), and C_8 (-▲-): number of clusters (J) = 8; number of sites (K) = 4; effect size variability (σ^2) = 0.01 (A_8), 0.1 (B_8), or 0.2 (C_8); reliability (rel) = 0.9 (A8), 0.7 (B_8), or 0.5 (C_8) and the proportion of explained variance by the blocking variable (B) = 0.15 (A_8), 0.10 (B_8), or 0 (C_8). Computed minimum detectable effect size (MDES) values associated with A_8, B_8 and C_8 are 1.49, 1.86, and 2.33, respectively.

educational effect. Such a value cannot even be considered "small" because it does not meet the minimum value (MDES) required for the study to maintain adequate statistical power. Researchers are not able to have confidence in the results if the MDES value is not sufficiently large, and it cannot be sufficiently large unless the variability within the system is controlled. Thus, once a district decides to conduct research using fewer than 10 schools, researchers must accept the limitations of the statistical parameters associated with the physical sci-

ences system: variability must be limited and the MDES value must be significantly larger than those typically reported using the social sciences statistical parameters.

Using a study design similar to that used in the physical sciences is an effective approach to using fewer schools in randomized controlled trials while maintaining adequate statistical power. This allows schools to manage the costs and logistics associated with completing a study. However, to be useful, theory must translate into practice. The utility of knowing how to use theoretical information to develop school-randomized controlled trials can be illustrated using the reported number of discipline referrals in schools, a measurable outcome of interest for both character education and prevention programs.

Stratification of Schools by Level of Variance

Archival discipline referral data was obtained from 18 high schools located in New York or California that were interested in being involved in a randomized controlled trial investigating the effectiveness of a school-wide character education/prevention program. However, because of financial and logistical constraints, only 8 schools could be included in the study. Of the original 18 schools that applied, 11 schools were able to provide stable prestudy data. The other schools either had definite increasing or decreasing trends in the 3-year data or were undergoing major changes in the upcoming academic year (i.e., grade levels being split between the existing school and a newly built school) so that measuring outcomes would be very difficult.

Once schools with stable data were identified, we considered how best to determine which of the 11 schools would be accepted into the study. As can be seen in Table 1, the percent variability in the discipline referral data ranged from 1% to 42% in the schools that were able to provide stable data. Note that the variability in the data for 8 of the schools was 16% or less. The percent variability for 3 of the schools was 37% or greater. MDES values are expressed in terms of standard deviation (SD) units (Schochet, 2005); thus, the percent variation represents the percent decrease in discipline referrals that must be met in order to achieve 1 SD unit or an effect size approximately equal to 1. For example, in school #1, the intervention under investigation would need to produce a 13% decrease in discipline referrals in order to achieve an effect size equal to 1 (assuming no change in the control schools). In school #11, the same intervention would need to produce a 42% decrease in discipline referrals to achieve an equivalent effect size. The power curves presented in Figure 2 demonstrate that, when using 8 schools (clusters), an effect size of at least 1.495 *SD* units (obtained from the "best case" scenario power curve, A_8) must be met in order to maintain adequate power. Therefore, to achieve an effect size of 1.495 *SD* units so that researchers can have reasonable confidence in the results, the intervention under investigation would need to produce at least a 19% decrease in discipline referrals in school #1 and at least a 63% decrease in school #11.

Information from Table 1 and Figure 2 was used to determine whether or not an 8-school study would be feasible given the expected outcomes that are obtained as a result of implementing the intervention. Previous case studies (The Flippen Group, 2006) employing quasi-experimental designs have demonstrated that the school-wide intervention being evaluated can reduce discipline referrals in schools 25% to 75% (depending on a number of variables including the commitment of school leadership and the fidelity of implementation). The choice of schools becomes clear if school researchers want to have the ability to measure smaller (25%) reductions in discipline referrals: schools with standard deviations representing less than 25% variability in the 3-year mean for discipline referrals are the best candidates for the study. Therefore, we accepted into the study the 8 schools whose discipline referral data fell within the < 25% variability stratum (see Table 1).

TABLE 1
Stratification of SCHOOLS on the Basis of Prestudy Discipline Referral Data

Condition	*School #*	*Mean Campus Enrollment*	***Mean # Discipline Referrals (± SD)***	*SD/Mean # Discipline Referrals (expressed as %)*	*Variability Stratum*
Intervention	1	112	**200** (± 25)	13%	
Control	2	347	**185** (± 21)	11%	
Control	3	540	**785** (± 20)	3%	
Intervention	4	373	**1,739** (± 149)	9%	< 25%
Intervention	5	429	**647** (± 75)	12%	
Control	6	1,248	**1,239** (± 13)	1%	
Intervention	7	2,106	**8,236** (± 1318)	16%	
Control	8	3,195	**8,378** (± 1214)	14%	
	9	2,350	2,692 (± 985)	37%	25-40%
	10	1,594	1,556 (± 635)	41%	
	11	2,834	4,028 (± 1707)	42%	> 40%

Discipline Referral Outcomes

Intervention schools received the program training during the 2008-2009 school year whereas the control schools did not receive the training or any training similar to it. At the end of the school year, all principals reported the schools' behavioral and academic data. The discipline referral numbers were converted to standard deviation units and were analyzed by univariate ANOVA. Graphic representation of the change from the mean ($\mu = 0$) in terms of standard deviation units (z-score) for each school is provided in Figure 3. The effect size, reported as the standardized mean difference (Hedge's g) was computed to be -2.1 (corresponding to a 22% reduction in the mean number of discipline referrals in the intervention schools and an 11% increase in the mean number of referrals in the control schools). A *p* - value of 0.0189 (less than the critical alpha level of 0.05) was computed from the F-statistic ($F_{0.05}$ [1,6] = 10.165). An effect size of $\left|-2.1\right|$ (decrease in discipline referrals) is in the MDES range ($\left|1.49 \text{ to } 2.33\right|$) that was originally determined (in Figure 2) to be required to achieve adequate power. The power of the present study was computed to be 0.78.

DISCUSSION

These results demonstrate that when district leadership teams are considering a school-randomized, controlled trial to determine the effectiveness of a comprehensive intervention, they can use power analyses to determine whether or not an appropriate study design can be developed such that changes in outcomes can be measured if they occur, thus providing justification for expending resources to conduct the study. Using a method to stratify schools based upon the variability in preimplementation outcome scores and then limiting the amount of variability allowed into the study prior to randomization of schools allows researchers to incorporate critical design elements that can decrease costs by increasing study precision such that fewer schools are needed to meet statistical requirements.

The argument can be made that including in a research study only those schools with stable baseline data for a particular outcome (or outcomes) is controlling the prestudy data such that the results cannot be replicated in schools with great variability in the same outcome measure(s). However, controlling the variability in the prestudy data in cluster-randomized educational trials is analogous to chemists pur-

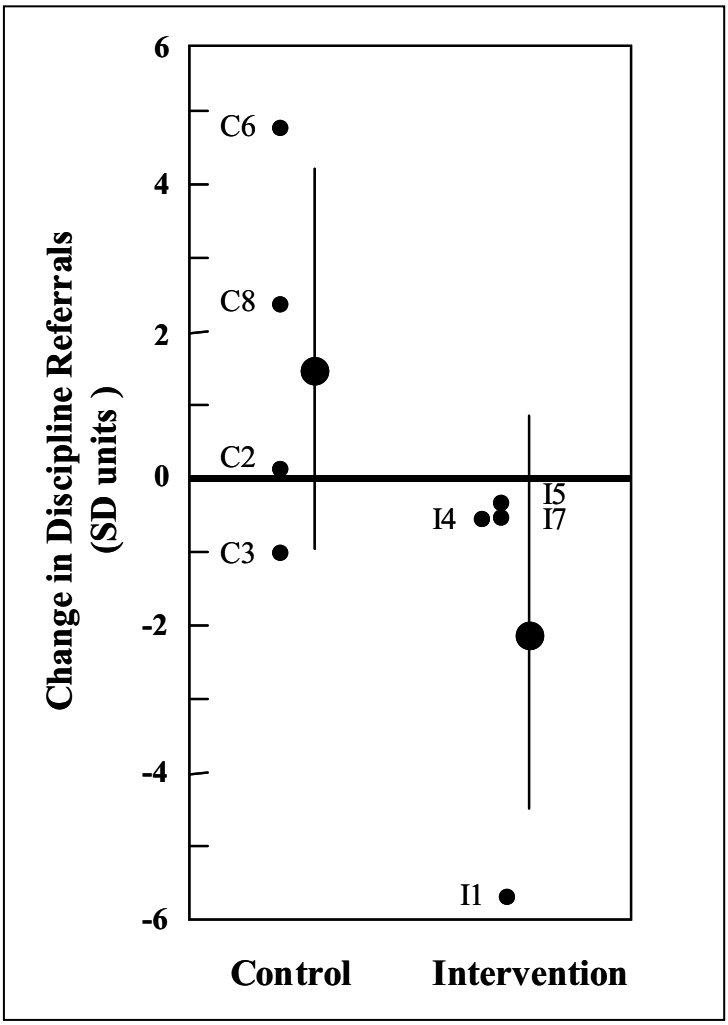

FIGURE 3

Changes in discipline referrals using standard deviation units (z-values). The small points represent individual schools; large points with error bars represent the mean change for each group of four schools (control vs. intervention). Positive values indicate increases in discipline referrals whereas negative values indicate decreases.

chasing test equipment (i.e., 96-well plates) whose between-well variability is tightly controlled. Because schools are randomly assigned to intervention and control groups after they are admitted into the study, and because random assignment is considered to be the best method to eliminate selection bias such that strong evidence of program effectiveness is obtained, the results of a study in which researchers control for variability is replicable, but if schools with high variability in the prestudy data are admitted into the study, many more schools (or clusters, *J*) would be needed to overcome the higher level of variability. As in other areas of research (such as in the medical field), small and large trials incorporate differing levels of variability and serve distinct purposes in social research strategies.

Small trials contain costs while demonstrating the feasibility of the research design and the effectiveness of the intervention in a limited number of participants. Large trials cost more, but provide evidence of the effectiveness of the intervention in a more diverse population of participants.

In an article relating statistical precision to math and reading outcomes in samples of 20, 40, and 60 schools, Bloom et al. (2007) ask the question—How small must the minimum detectable effect size be for an educational evaluation? The researchers comment that determining the required precision needed for educational interventions is in a "state of flux" because program effects as small as 0.1-0.2 may produce significant changes in student achievement (Kane, 2004; Nye, Hedges, & Konstantopoulos 1999) and thus may be relevant to policymakers. This level is lower than that of Cohen's (1988) widely-used designation of 0.2 for a small significant effect size, but is in line with Lipsey's (1990) meta-analytic categorization of effect sizes with values between 0 and 0.32 as the lower distribution for small, but significant, effect sizes.

The question of how small the change in an educational outcome can be (i.e. the effect size) and still have relevance for policymakers is a different question that that asked in the present study. In this report involving the statistical requirements for studies in which fewer than 10 clusters can participate, the question asked is not how small the effect size can be and still be educationally important, but rather, how large it must be in order for researchers to achieve adequate power so that they can have confidence in the outcomes obtained in educational trials using fewer schools. The results obtained from the data from 8 public high schools demonstrate that researchers have the ability to determine in advance whether or not a research study is likely to have the statistical power to detect non-negligible outcomes if they occur.

Stratification prior to randomization from matched pairs can help school districts avoid spending money for a study that cannot effectively measure the outcomes of a behavioral intervention such as a comprehensive character education program. Using detailed baseline data prior to a study can provide school researchers with the statistical precision that they need to use fewer schools per study, thus lowering the cost and logistical difficulty of the study. Additionally, obtaining this data can provide potential funding agencies and other stakeholders with concrete evidence demonstrating the feasibility of a statistically relevant outcome, assuming that the intervention is implemented fully and with fidelity.

CONCLUSIONS

The No Child Left Behind Act requires that evidence-based programs be used when applying for certain types of federal funding. Random assignment of schools to intervention or control groups is used to eliminate selection bias, thus providing the strongest evidence of the true impact of the intervention. However, obtaining evidence-based research in this manner when implementing school-wide interventions may seem too difficult for districts to achieve for budgetary or logistical reasons. For either reason, practitioners may choose not to engage in high-quality research to determine whether an innovation works, but may instead choose to develop ways to recycle current knowledge (Viadero, 2009).

Deciding to undertake rigorous research involving character education or prevention interventions does not need to be an "either/or" scenario when it comes to interventions requiring cluster-randomized study designs. Researchers can develop high-quality research studies if they acquire specific pre-study data on schools, and then select schools with stable baseline data with low prestudy variability associated with one or more desired outcomes. Doing so limits the variability that impacts effect size (δ), thus increasing acceptable power (0.8) with fewer schools (J) and lower cost.

Acknowledgment: The author is grateful for invaluable research discussions with Drs. Thomas Lickona, Matt Davidson, and Vlad Khmelkov (SUNY-Cortland); Vic Battistich (late) and Marvin Berkowitz (U. Missouri-St. Louis); Jessaca Spybrook (U. Western Michigan); and Victor Willson (Texas A&M University).

REFERENCES

Aber, J. L., Brown, J. L., & Jones, S. M. (2003). Developmental trajectories toward violence in middle childhood: Course, demographic differences, and response to school-based intervention. *Developmental Psychology*, *39*(2), 324-348.

Battistich, V., Schaps, E., Watson, D., Solomon, D., & Lewis, C. (2000). Effects of the child development project on students' drug use and other problem behaviors. *Journal of Primary Prevention*, *21*(1), 75-99.

Battistich, V. (2003). Effects of a school-based program to enhance pro-social development on children's peer relations and social adjustment. *Journal of Research in Character Education, 1* (1), 1-16.

Berkowitz, M., & Bier, M. (2006a). *What works in character education: A research-driven guide for educators.* Washington DC: Character Education Partnership. Retrieved from http://www.characterandcitizenship.org/research/wwceforpractitioners.pdf

Berkowitz, M., & Bier, M. (2006b). *What works in character education: A report for policy makers and opinion leaders.* Washington DC: Character Education Partnership. Retrieved from http://www.characterandcitizenship.org/research/WWCEforpolicymakers.pdf

Best Evidence Encyclopedia (BEE): Empowering Educators with Evidence on Proven Programs (2011). *Interpreting effect sizes.* Baltimore, MD: Johns Hopkins University Center for Data-Driven Reform in Education. Retrieved from http://www.bestevidence.org/aboutbee.htm

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Center for Disease Control. (2010). Youth violence: Best practices of youth violence prevention—A Sourcebook for community action. Retrieved from http://www.cdc.gov/violenceprevention/pub/YV_bestpractices.html

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Elias, M. J., & Arnold, H. (2006). *The educator's guide to emotional intelligence and academic achievement: Social-emotional learning in the classroom.* Thousand Oak, CA: Corwin Press.

Flay, B. R., Allred, C. G., & Ordway, N. (2001). Effects of the Positive Action program on achievement and discipline: Two matched-control comparisons. *Prevention Science, 2*(2), 71-89.

Greenberg, M. T., Kusché, C. A., Cook, E. T., & Quamma, J. P. (1995). Promoting emotional competence in school-aged deaf children: The effects of the PATHS curriculum. *Development and Psychopathology*, *7*, 117-136.

Holtzapple, C. K., Buckley, S. A., & Stanker, L. H. (1997). Production and characterization of monoclonal antibodies against sarafloxacin and cross-reactivity studies of related fluoroquinolones. *Journal of Agricultural and Food Chemistry, 45*(5), 1984-1990.

Holtzapple, C. K., Pishko, E. J., & Stanker, L. H. (2000). Separation and quantification of two fluoroquinolones in serum by on-line high-performance immunoaffinity chromatography. *Analytical Chemistry, 72*(17), 4148-4153.

Holtzapple, C. K. (2009). *Controlling for baseline covariates to improve the statistical power of efficacy trials involving whole school prevention programs.* 17th Annual Meeting of the Society for Prevention Research. Retrieved from http://www.preventionscience.org/SPR_09_program_COMPLETE.pdf (abstract #224, p. 63).

Kane, T. (2004). The impact of after-school programs: Interpreting the results of four recent evaluations. New York: William T. Grant Foundation.

Lickona, T., & Davidson, M. (2005). *Smart & good high schools: Integrating excellence and ethics for success in school, work and beyond.* Cortland, NY: Center for the 4th and 5th Rs (Respect & Responsibility). Washington, DC: Character Education Partnership.

NIDA Notes (2002). *Risk and Protective Factors in Substance Abuse Prevention, 16*(6). Retrieved

from http://www.drugabuse.gov/NIDA_Notes/NNVol16N6/Risk.html

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: SAGE.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee Class Size Experiment. *Educational Evaluation and Policy Analysis, 21,* 127-142.

Office of Juvenile Justice and Delinquency Prevention (2004). *OJJDP Model Programs Guide, version 3.0.* Retrieved from http://www2.dsgonline.com/mpg/prevention.aspx?continuum=prevention

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multi-site randomized trials. *Psychological Methods, 5*, 199-213.

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2004). Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design software. University of Michigan.

Raudenbush, S.W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5-29.

Schochet, P. (2005). *Statistical power for random assignment evaluations of education programs.* Princeton, NJ: Mathematica Policy Research.

SPSS (2010). Software versions retrieved from http://www.spss.com/software_version/index.cfm?product=base/

Spybrook, J., Raudenbush, S.W., Liu, X., Congdon, R. (2006). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" (version 1.76) software.

Spybrook, J. (2008). Experimental designs and statistical power of group randomized trials funded by the Institute of Education Sciences. Society for Research in Educational Effectiveness (SREE) 2008 Annual Conference. Retrieved from http://www.meetinglink.org/educationaleffectiveness/2008/conference/submission/abstracts/jessacah20071024135114/SREEAbstractTemplate.doc

Spybrook, J., Raudenbush, S. W., Liu, X., Congdon, R. (2009). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" (version 2.0) software. Retrieved from http://www.wtgrantfoundation.org/resources/overview/research_tools/research_tools

The Flippen Group. (2006). *Report to the Texas Education Agency*. Retrieved from http://www.flippengroup.com/pdf/funding/TEA8CaseStudies.pdf

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes.* Washington, DC: What Works Clearinghouse. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/essig.pdf

Viadero, D. (2009, Jan. 27). "Scientifically based" giving way to "development," "innovation." *Education Week* [Online Issue]. Retrieved from www.edweek.org/ew/articles/2009/01/28/19rd_ep.h28.html?print=1

W.T. Grant Foundation. (2011). Optimal Design Software Vesion 2.0. Retrieved from http://www.wtgrantfoundation.org/resources/overview/research_tools/research_tools

What Works Clearinghouse. (2007). Character Education Overview. Retrieved from http://ies.ed.gov/ncee/wwc/reports/character_education/topic/

What Works Clearinghouse. (2008). WWC Procedures and Standards Handbook (version 2.0, December 2008). Retrieved October 16, 2009 http://ies.ed.gov/ncee/wwc/references/idocviewer/Doc.aspx?docId=19&tocId=8

Zins, J. E., Elias, M. J., Greenberg, M. T., & Weissberg, R. P. (2000). Promoting social and emotional competence in children. In K. M. Minke & G. C. Bear (Eds.), *Preventing school problems—promoting school success: Strategies and programs that work* (pp. 71-100). Bethesda, MD: National Association of School Psychologists.