

The effects of K-12 school enrollment interventions in developing nations

**3ie Synthetic Reviews – SR004
Protocol
November 2009**

Review team:

Anthony Petrosino, Ph.D. (Contact Author)
Senior Research Associate
Learning Innovations at WestEd
200 Unicorn Park Drive, 4th Floor
Woburn, Massachusetts 01801
781-481-1117
781-481-1120
apetros@wested.org

Claire Morgan, M.A.
Research Associate
Learning Innovations at WestEd
200 Unicorn Park Drive, 4th Floor
Woburn, Massachusetts 01801
781-481-1119
781-481-1120
cmorgan@wested.org

Robert F. Boruch, Ph.D.
University Trustee Chair Professor
Graduate School of Education and
Statistics Department, Wharton School
University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104
215-898-0409
robertb@gse.upenn.edu

Funding: 3IE

1. Background

It has been well established that education is key to economic development and social welfare. Investments in education yield returns in poverty reduction, improved health outcomes, and economic growth (UNESCO, 2007; Hannum & Buchmann, 2004; Herz & Sperling, 2003). In addition, increased access to education contributes to increased political participation and more equitable sharing of economic and political power (Birdsall, 1999). Education for girls is particularly critical, as improvements in the infant mortality rate, child nutrition, and school enrollment are closely associated with maternal education (Birdsall, Levine, & Ibrahim, 2005; Herz & Sperling, 2003; World Bank, 2008). Yet, more than 100 million primary school aged children are not in school, and of those that are, many—49 percent in Africa, for example—do not complete primary school (Birdsall, Levine, & Ibrahim, 2005).

Low educational attainment—or the inability of students to complete their primary and secondary school education—in the developing world is the combined result of children who do not enroll, children who do not progress, and children who drop out (World Bank, 2004). Children may not enroll or complete their schooling for a number of reasons. Research indicates that there are economic reasons and other structural forces that present barriers. For example, in some countries such as India, Mali, and Burkina Faso, school enrollment is very low, due to issues such as the cost of schooling (both direct and opportunity costs), poor school infrastructure, teacher shortages, and safety and sanitation problems (Birdsall, Levine, & Ibrahim, 2005). In others, such as many Latin American countries, enrollment may be nearly universal, but retention and completion may be quite low (*ibid*) for a myriad of reasons, including those mentioned above, as well as poor health of students or members of their households (Glewwe & Miguel, 2008; UNESCO, 2007), teacher absenteeism or malfeasance (World Bank, 2004), and curricula that do not match students' needs (Glewwe, Kremer, & Moulin, forthcoming). Value systems held within the country may also diminish the importance of enrolling children in school (e.g., Brembeck, 1962).

Furthermore, developing nations face significant enrollment and completion disparities between segments of the population such as rich and poor, boys and girls, urban and rural dwellers, and combinations of these factors (Birdsall, Levine, & Ibrahim, 2005). For example, in India, the gender gap (favoring boys) for children from the richest households is only 2.5 percent, while disparity for children from the poorest households is 24 percent (Filmer, 1999 as cited in Birdsall, Levine, & Ibrahim 2005). In many African nations, rural rates of enrollment lag far behind the very modest national rates, particularly for rural girls, whose rate of enrollment is less than 15 percent in several countries (*ibid*). In addition, ethno-linguistic diversity, disabilities, and conflict situations in fragile states create further barriers for school participation in developing nations (Birdsall, Levine, & Ibrahim 2005).

In light of compelling evidence that links expanded education systems and socioeconomic development while highlighting the importance of policies to offset inequality in access (UNESCO, 2007), and spurred by the donor community and such initiatives as the Millennium Development Goals and Education for All, governments in developing nations are, to varying degrees, making efforts to increase school enrollment and equity. Building new schools to increase ease of access in remote areas is one intervention used in developing nations (e.g., Filmer, 2004). Other efforts include improving school infrastructure and safety and abolishing school fees, as well as implementing targeted policies to reach the most marginalized children.

Such policies include school feeding programs, flexible schooling models for working children, school-based health interventions, and various types of financial subsidies and conditional cash transfer systems. For example, several Latin American governments and non-governmental partners have experimented with programs that transfer money directly to disadvantaged households—such as in rural, indigenous, migrant, or slum communities—in exchange for children’s school enrollment and attendance (UNESCO, 2007). In Asia, stipend programs encourage the transition of girls to secondary school (ibid).

Evaluations of some of these recent policies and programs to increase school enrollment and persistence in developing nations include a number of randomized field trials and rigorous quasi-experimental studies. Randomized experiments evaluating conditional cash transfer programs in Latin America include the seminal *Progresá* school subsidies experiment in Mexico, which gave educational grants to poor mothers of children enrolled in school with good attendance. Communities were randomly assigned to intervention or control conditions and positive impacts for school enrollment and other factors were demonstrated (Schultz, 2004). Similarly, in Ecuador, a lottery provided cash vouchers to randomly selected families in exchange for enrolling their children in school; control families were placed in a “wait-list” condition until the study was completed. The early results were positive, increasing school enrollment by 10 percent and reducing child labor by 17 percent (Lopez-Calva, 2008). In addition, Filmer & Schady (2006) found that a scholarship program for girls in Cambodia making the transition from primary to secondary school had a large, positive effect on enrollment and attendance.

Randomized trials of school-based health interventions include a school feeding program in rural Peru, in which schools were randomized to implement a high-quality, ready-to-eat breakfast program or to a control group, with positive results for school enrollment and other outcomes (Cueto & Chinen, 2008). Glewwe & Miguel (2008) review randomized evaluations of school-based health interventions such as that of Miguel and Kremer (2004), which found that absenteeism in Kenyan schools in which students received deworming treatment was 25 percent lower than in comparison schools, and that deworming increased schooling by 0.14 years. Recent randomized evaluations of other types of programs aimed at increasing enrollment and completion include that of Glewwe, Kremer, and Moulin (2007), who found that providing textbooks to students in randomly selected rural primary schools in Kenya had no effect on dropout or repetition rate. The Millennium Challenge Corporation funded a regression discontinuity study to assess the impact of school construction and other associated interventions on female student school enrollment in 132 communities in Burkina Faso versus 161 communities not selected for ‘treatment’, finding positive results on enrollment (Levy et al. 2009).

To our knowledge, a systematic review of randomized controlled trials and quasi-experiments of school enrollment and persistence strategies in developing nations has not yet been reported. By systematically gathering and analyzing rigorous research about the program effects of primary and secondary school enrollment and secondary school completion policies, our review will provide key evidence to inform the next wave of development efforts in this area.

2. Objectives

For this project, we will be collecting studies that respond to the question: *What are the documented impacts of school enrollment policies and programs in developing nations (see definition below) on enrollment and persistence outcomes (e.g., primary and secondary school enrollment, attendance, retention, primary to secondary transition, and secondary school completion), and on learning outcomes (e.g., test scores, grades, etc.)?*

3. Methodology

Criteria for inclusion and exclusion of studies in the review

For this project, we will only include those studies that have the following characteristics:

(1) Randomized controlled trials or quasi-experimental evaluations with evidence of equating. Our review includes evaluations that randomly assign entities (at any level) to intervention or control conditions. In the studies we have examined to date, the control or comparison group (the counterfactual) is either a condition that receives no intervention or receives the usual or standard practice. Only well-implemented randomized experiments can provide statistically unbiased estimates of an intervention's effect (Boruch, 1997). Although there are comparatively very few randomized experiments in developing nations compared to other types of evaluation studies (Glewwe and Kremer, 2006), there has been a "randomized revolution" in development work in recent years (e.g., Duflo and Kremer, 2005). However, because randomization is not possible in certain evaluation scenarios (e.g., in retrospective evaluations or when the program has already been implemented), we will also include evaluative studies that use quasi-experimental designs, provided that they offer evidence that the comparison groups were equated. Such quasi-experiments include regression discontinuity design, propensity scores, covariate matching, and other pre or post intervention matching. Although quasi-experiments cannot control for unknown or unmeasured confounding factors, there is conflicting literature on whether the estimates from such designs approximate those from randomized experiments (e.g., Oliver et al 2008). We include both types of studies and will include study design as a moderator in our later analyses (see discussion on moderator variable analysis in later section).

(2) The evaluations have to have taken place in a country classified as a "low or middle income nation" by the World Bank at the time the intervention being studied was implemented. This is the definition used by 3ie. The World Bank determines low and middle income nation status by calculating *the Gross National Income (GNI) per capita*; this is essentially the average citizen's income. As of 2008, 151 nations are included in these categories. The categories are: low income (\$975 or less); lower middle income (\$976-\$3,855); and upper middle income (\$3,856-\$11,905). Low and middle income nations are often referred to as "developing economies" and overlap considerably with the United Nations listing of "developing nations."

(3) The evaluations have to assess the impact of a K-12 school enrollment strategy. Programs designed to boost preschool, college, or university enrollment, or other strategies, will not be included.

(4) *The evaluations have to include at least one outcome measure of school enrollment or persistence.* Other measures, including other learning outcomes (e.g., grades, test scores) or impacts on behavior, health, child labor, costs, equity, or attitudes or satisfaction levels will also be collected provided the study includes at least one measure of school enrollment/persistence or learning.

(5) *The evaluation study report is published or available through December 2009, without regard to language or publication type.* We will also search for trials published up to and including December 2009, without regard to the start date of publication. Studies published or available after 2009 will not be eligible. Citations from any source that fall within those dates, whether databases, websites, in reference sections of available documents, or sent to us by colleagues, will be included. However, we anticipate that most eligible studies will have been published after 1960. In concert with Campbell principles, we will attempt to find English and non-English studies. In addition, we will include published and unpublished studies (e.g., from conference papers, dissertations, technical reports).

Example of studies that would be included in our review

The Burkina Faso evaluation conducted by Mathematica for the Millennium Challenge Corporation (Levy, et al. 2009) is an example of an evaluation that would be included in our review. Using regression discontinuity design, the study compares the results, at the village level, on girls' school enrollment for those villages that received new school construction and other interventions (via the BRIGHT program) versus those that did not. Results indicate that girls' enrollment 16-19 percent in the experimental villages. The background section included references to other examples of evaluations that would be eligible for our review. For example, the Ecuador lottery study that randomly assigned families to receive cash vouchers in exchange for enrolling their children in school would be included (Lopez-Calva, 2008).

Example of studies that would not be included in our review

One study that would not be included in our review is that of Bobonis, Miguel, and Sharma (2006), who evaluated a health program that provided iron supplementation and deworming medicine to preschool age children in poor urban areas of Delhi. Although this was a randomized evaluation in a developing nation that reported effects on school absenteeism, we would not include it in our review because it does not evaluate a *K-12 strategy*, but focuses on the preschool population. Another example of a study that would not be included in our review was conducted by Lockheed and her colleagues (1986) in Thailand. They studied the impact of providing textbooks to Thai students. Although there was a question of whether they used a truly equated comparison group, the study was nonetheless excluded because it did not include any outcome of enrolment, attendance, dropout, or persistence, but exclusively focused on academic achievement.

Search strategy for identification of relevant studies

Our goal is to identify both published and grey literature. Many of the databases in (1) below include grey literature (e.g., ERIC). The British Library indexes conference proceedings and makes these available in its "Integrated Catalogue." Web searches will often identify literature that is made available at websites but not published in

journals. Our contact with colleagues is designed to get at more of the grey literature. To accomplish our goal, we will use five major strategies:

- (1) *Electronic searches of bibliographic databases.* Researchers will use available online resources and databases at WestEd, the University of Pennsylvania, Boston Public Library, and the University of Massachusetts, including ERIC, British Education Index, PAIS International/Archive, Sociological Abstracts and World Bank Documents. See Appendix A for additional databases that will be searched. We will also use the “advanced search” options in Google for broad searches of the World Wide Web.
- (2) *Hand searches of relevant journals.* Because electronic searches often miss relevant studies, we will hand search the table of contents, and the abstracts when necessary, of all issues of the journals most likely to publish studies on this topic. From our initial searches, the five journals that we have identified for hand search are: *Economic Development and Cultural Change*, *International Journal of Educational Development*, *Journal of Development Economics*, *World Bank Research Observer*, and the *World Bank Economic Review*.
- (3) *Specific examinations of online holdings of international development organizations and research firms.* This would also include international or national/federal agencies that either conduct or would be aware of possibly relevant evaluations in developing nations, including the World Health Organization, the U.K. Department for International Development (DfID), and the United States Agency for International Development (US-AID). It would also provide coverage of websites with great relevance to international development, including the Network for Policy Research, Review and Advice on Education and Training (www.norrag.org). Research firms such as RTI International (particularly its international education division at http://www.rti.org/page.cfm/International_Education) and the Academy for Educational Development (www.aed.org) will also be part of this search strategy.
- (4) *Citation chasing.* The reference section of every retrieved report will also be checked to determine whether any possible eligible evaluations are listed. As noted in the eligibility criteria, we are not exclusively seeking English language reports. We will ask our colleagues from other nations for help in identifying any non-English studies. WestEd also has employees bilingual in Spanish, French, Japanese, and Chinese who can translate abstracts or full-text documents in non-English to determine their eligibility for this review.
- (5) *Contacting the ‘informal college’ of researchers in this area.* There is a network of researchers that are conducting or are aware of experimental and quasi-experimental studies relevant to developing nations. We will identify the lead authors of such studies or relevant documents (e.g., reviews, non-evaluative studies), identify their emails from a Google search of the World Wide Web, and email them query letters. A full listing of all persons contacted will be included in the final review.

Keyword strategies for bibliographic databases

The databases in Appendix A can be somewhat idiosyncratic. Thus, we believe the best strategy is to conduct a broad search of the available databases that errs on the side of sensitivity rather than specificity. In other words, we would rather get many

titles and abstracts to sift through rather than potentially miss relevant citations because our search terms were drawn too narrowly.

We will use two different search strategies for these databases, depending on their focus. If the database is focused on education (such as *ERIC*), we will use broad searches that identify evaluation studies conducted in developing nations. To do this, we will use the following keywords to find experimental and quasi-experimental outcome studies: "random," "experiment," "control," "evaluate," "trial," "impact," "effect," "comparison group," "match," "discontinuity," "propensity" and "outcome." Our plan is to use truncated versions of these words; for example, we will use "experiment*" to capture titles and abstracts with the word "experiment," "experimental," "experimenting," "experimenter," and "experiments."

Second, we will combine those keywords with ones that focus the search on developing nations, such as the use of terms like "developing," "third world," and "impoverished" with "nation," "country," or "region." In addition, we will use the names of specific developing nations, such as India, Mexico, etc., and the names of regions, such as Africa, Latin America, and Asia. If there is a geographic descriptor for country or region, we will incorporate that into our search process. We will also include keywords that focus on the outcomes of interest including dropout, enrollment, enrolment, and attendance. Such searching is an iterative process and we will modify as we retrieve studies. This strategy may produce a number of false positives, but our experience is that examining the abstracts is not time consuming and researchers can go through them quite quickly.

Wherever possible, we will limit our searches by a descriptor that indicates the grades that our review is targeting: K-12 (for example, selecting primary and secondary education, or elementary, middle, and high schools). This would have the advantage of screening out preschool and college age studies.

If the database is not focused on education (e.g., *Sociological Abstracts*), the above strategy must then be supplemented by something that identifies educational research. In some databases, that will be a classification code; for example, in *Sociological Abstracts (Sociofile)*, one can limit the abstracts to those dealing with "sociology of education". But in many of others, there is no classification code. Whether classification codes exist or not, we will use truncated versions of keywords related to the educational outcomes of interest, such as dropout, attendance, and enrollment/enrolment to try to reduce the number of false positives. Again, if there are geographic descriptors for country or region, we will incorporate that into our search process.

We recognize that specific search strategies may have to be developed for each database. What works in identifying potential studies in ERIC will not work in searching World Bank documents. The appendices to our final review report will carefully document all keywords used for each database to permit replication.

Retrieving and Final Screening of Studies

Search methods will identify a large number of citations and abstracts. Many of these will be easily excluded as not being relevant to the proposed review. In some cases, however, they will identify potentially eligible studies. The first and second authors will review all citations and determine if the cited study should proceed to a second screening, i.e., is a potentially relevant study. If so, the full text documents of those

potentially eligible studies will be retrieved and screened by the first two authors before the study can be formally included in the review. Fortunately, with the advent of the Internet and full-text electronic journal access, we will be able to rapidly retrieve the reports to do a more thorough reading. When a full text report is received, we will scan it to ensure that it includes randomization or quasi-experimental equating of study subjects and includes at least one outcome of school enrollment or persistence. If the first two Investigators do not agree on the inclusion of a particular study, it will be excluded and documented in the final report.

We have established a bibliographic reference database to maintain a log of all included and excluded studies. The log includes a field that allows the research team to document the reason for exclusion.

Extracting Information from Each Study

We have designed a preliminary instrument to guide us in recording information from each study (see Appendix B). Although the instrument contains several open-ended items, these will be collapsed when appropriate into a smaller number of categories to permit further analysis. For example, items such as "how equating was performed" can be collapsed into three or four larger categories representing the most frequent responses (e.g., discontinuity, covariate matching, propensity score, post-hoc statistical matching) and an "other" category that captures all those responses that do not fit into the most common methods of equating in this set of studies.

The instrument has items in the following areas:

Researcher, Study and Contextual Characteristics:

Study reports can be used to provide information about the publication and characteristics about the experiment and the context. For example, we will extract data about the type of publication the study was reported in and the setting in which the trial was conducted. If the documents provide information on the context in which the study takes place, we will also include it.

Study Methods and Methodological Quality:

We will extract information about the randomization, quasi-experimental assignment, and other methodological aspects of the evaluation. The level of assignment and whether the study included multiple analyses at different levels will also be coded. It is especially critical that information about three key issues in the implementation be extracted from each study report:

- a) *How the groups were equated and whether any problems with equating were reported.* The integrity of a randomized experiment or a quasi-experiment largely rests on how faithfully the equating procedures were implemented. We will code information about randomization and the quasi-experimental matching or equating procedures that were used in the study. In randomized experiments, this includes how much of the originally randomly assigned sample actually received the treatment (slippage from the 'intention to treat' sample). We will code this information using a two-stage process. The first stage is a more detailed gathering of the facts about the assignment. The second stage will be

comprised of ratings by two reviewers that will indicate the degree to which group equating was comprised by any reported problems.

- b) *Whether the researchers report a loss of participants from the initial assigned sample at the end of the study, how much attrition is reported, and whether the attrition differentially affects one group or the other.* Such attrition, if it is significant, can comprise the equating of groups, particularly if different types of people drop out from the intervention than dropped out from the other conditions. We will code specific information on the amount of attrition (if it occurred) and whether it was differential in nature. We will conduct sensitivity analyses to examine if results change based on dropping studies that experienced different levels of attrition (10%-25%; 25%-50%; 50% or more), or if results change if studies that report differential attrition (significant losses from one group or the other, with a difference in attrition of 5% or more between the groups). These are obviously subjective classifications, but the goal is to determine if the attrition comprised the study findings.
- c) *Whether the program experienced significant implementation and fidelity problems.* The first two issues deal with the implementation of the evaluation. This issue deals with the implementation of the program; there may be no observable program impact because no 'real program' was ever implemented. We propose two-stage coding of implementation. First, we will code, in descriptive and qualitative form, any implementation problems noted by the investigators. Second, we will then rate the degree of implementation problems (with the standard being how the implementation problem affects a 'fair test' of the program under investigation) as 'high,' 'moderate,' or 'low.'

Intervention and Control Conditions data:

These items will solicit detailed descriptions of the intervention and control condition, including the 'dosage' of the treatment being implemented, and the number of participants assigned to each group. We anticipate that the evaluations in this review sample will be comprised of a single intervention and a single control group. When this is not the case, we will select the most policy relevant groups to compute our experimental versus control condition contrast. In most cases, it will be the groups that experience the greatest contrast between conditions, i.e., the most intensive intervention condition versus the least intensive control condition. We recognize the importance of documenting these decisions for full transparency.

Participants in the Trial data:

These items solicit detail about the type of participants in the trials, including information on the country where the study took place, the nationality of the participants, the age and school level targeted, gender, whether an urban or rural setting was involved, and the socioeconomic status of the students.

Outcome data:

For each eligible study (each eligible study will have, at minimum, one outcome measure of enrollment or persistence), we will extract information on reported outcomes including impacts on learning, health, child labor, costs, and equity. We will also code any other outputs or data on key 'mechanisms' that would provide

clues as to why the intervention did or did not have its intended impact.

Handling multiple reports on the same experiment

Note that investigators may publish several articles on the same study. Our unit of analysis is the individual evaluation and not the individual research article, and so it is reasonable to extract information from all documents to complete the coding instrument for one experiment. When reports on the same study contain conflicting information, we will employ a number of strategies, including contacting the original investigator(s) for resolution.

Criteria for determination of independent findings

Each study will be represented by a *single effect size* to prevent the analysis from being compromised by non-independence (multiple effect sizes from one study). Although some evaluations may report just a single outcome at one time interval, it is more likely that evaluation reports will include analyses at various time intervals and may use various constructs that reflect school enrollment and persistence. Therefore, decisions have to be made about what outcome will represent the effect size for that study.

| | Enrollment | Attendance | Dropout | Test Scores | Grades |
|----------------|------------|------------|---------|-------------|--------|
| First Effect | | | | | |
| Middle Effect | | | | | |
| Longest Effect | | | | | |

For this review, we will keep outcomes distinct. That is, we will analyze enrollment, school attendance, dropout, and other learning or non-educational (health, behavioral, etc.) outcomes separately. The Table provides an example of how four such outcomes would be reported. We do not know as of yet how such outcomes will be reported, i.e., will they be prevalence measures (percentage of groups that enroll or attend) or incidence measures (the mean rate for some outcome of interest, such as the mean number of days attended per student). If results are varied and include prevalence and incidence rates, we will discuss the best way to report these (combine or separate out) and make such decisions explicit in our review.

We also propose to report three different analyses to handle the studies that report outcome data at various time intervals. As the Table indicates, we will report effect sizes at first follow-up (the first time interval reported), the middle effect (the middle time interval closest to the exact point between the first and longest), and the longest effect (the effect size for the longest follow-up period). If one time interval (e.g., 1 year) is reported in the study, it will be used in all three analyses. If two time intervals (e.g., 6 months, 1 year) are reported, the results will be averaged and the mean will be reported for the 'middle effect.' If more than three time intervals are reported (e.g., 6, 12, 18 and 24 months), we will select the result that is closest to the exact middle. In this instance, the exact middle between 0-24 months is 12 months and would be reported as the middle.

If regression-adjusted estimates are reported for the experimental versus control groups, we will rely on them for any quantitative synthesis since they theoretically reduce statistical 'noise' that may have come from chance fluctuations or randomization violations (in the case of well implemented experiments) or uncontrolled variables (in the case of quasi-experiments).

Some studies report analyses at multiple levels (e.g., Schultz, 2004), i.e., for schools or localities and for studies. Our rule is to capture this information separately, but to compute effect sizes for the analysis done at the level of assignment. So, for example, in *Progresa*, the randomization was done at the locality/community level, and so the main effect size will be that computed for treatment and control localities. We will code information, however, about the analyses done at the student, family or school level.

Some studies also report effects at all grade levels (e.g., Schultz, 2004). This is very important to policy and practice decision-makers. The main effect will again be computed at the larger analysis level, so that if schools are assigned to groups, the effect size will be computed for all schools in treatment versus all schools in control. However, we will record subgroup effects such as breakdowns by grade and gender.

Details of study coding categories

To ensure that we achieve good coding reliability, we will have two of the co-authors read and record information from a random sample of reports (25%). We will assess coding reliability (i.e., inter-rater agreement) by using the percentage of agreement for each item, rather than reporting a global inter-rater reliability statistic. This will avoid inflating reliability measures with study characteristics that generally achieve perfect agreement (e.g. year of publication) with those that do not. Items with lower rates of agreement (less than 80%) will be investigated to determine the source for conflict. The authors will meet to resolve disagreements and discussing coded items. We will drop those items from our database in which resolution could not be reached, as well as items that lack clear interpretation.

Statistical procedures and conventions

The data will be entered into the Comprehensive Meta-Analysis (CMA), version 2. We will use CMA to statistically combine results from the evaluations.

We will report standardized mean differences (Cohen's d), as it is a very flexible effect size metric and many formulae are available to estimate Cohen's d from information often reported in evaluation articles (e.g., statistical test data, probability levels). Forest plots will be used to display the results from the effect sizes. The plot will display, for each study, the effect size, confidence intervals and significance level. The plot will also display the same for the average effect across studies. Note that this will be reported assuming a random effects model, and the estimate will be weighted by sample size. When describing results in the text, we will report the effect size, the confidence intervals and whether the analysis indicates that the result is statistically significant.

Because of the likely heterogeneity in interventions, samples, countries, and outcomes, we will assume random effects models in our analyses, which tends to be more conservative than the fixed effects approach.

For our analyses, we will conduct tests for heterogeneity to determine if the average effect size is a good representation of the sample of studies being used in the analysis. We anticipate the heterogeneity will be present, given the variations in intervention type, nation, sample populations and the like in these development studies. Using CMA, we will confirm heterogeneity in each summary analysis (of each

outcome at each of the three time intervals: first, middle and longest) through the Q-Value, which is reported as a summary indicator of the extent of variance across studies in the sample.

However, moderator analyses have to be approached carefully, as they are often based on small numbers of studies (the 'small cell' problem), and that such analyses can be significant by chance if large numbers of variables are considered (the 'capitalizing on chance' problem). We anticipate examining a small number of moderating variables as a source of heterogeneity, by comparing the effect sizes (assuming a random effects model) for the following potential moderators:

- *Randomized versus non-randomized studies.* An important question is whether the estimates from randomized experiments are different than those reported in studies that used non-random assignment. This review will examine the average effect size for experiments and compare it to the average effect size for quasi-experiments.
- *Intervention types.* An important policy question is whether developing nations and donor agencies are getting more 'bang for the buck' using one particular approach or another. We will recode the detailed intervention categories into discrete groups for analysis and compare the average effect sizes for these groups.
- *Gender.* Many interventions are specifically designed to increase female school enrollment. We will examine effect sizes for male, female and mixed gender samples.
- *Low versus middle income countries.* Our early searches have located eligible studies across the developing world, in approximately 40 different nations. However, many nations will be represented by a single study. To provide a more substantial analysis, we will examine effect size by World Bank classification of developing economies (low versus middle income nations).

We do not anticipate, at this time, conducting a study of publication bias. The reason is that economists conduct and report many of the studies relevant to this area. The tradition in the field of economics is to make unpublished papers available online. However, a large percentage of these are eventually published in economics journals. Therefore, papers may only be temporarily 'unpublished.' We are tracking all relevant documents for an evaluation, but do not believe a comparison of 'published' versus 'unpublished' studies would be particularly useful at this time.

Treatment of qualitative research

It is very unlikely, because of our focus on experiments and quasi-experiments, and our focus on quantifiable outcomes that can be converted into an effect size metric, that we will uncover much qualitative research. However, we will code the presence or absence of ancillary qualitative studies, what the studies focused on, and what the main findings are. Certainly, qualitative data from the experiments and quasi-experiments will be used to illuminate three particular areas: (1) the context for the intervention; (2) the theory or mechanisms by which the program is supposed to impact the ultimate outcomes; and (3) the quality and nature of the intervention and comparison condition.

Treatment of economic data

We will report on any economic data included in the primary studies that are included in the review. This includes information on the costs of the program, any analysis of the cost-effectiveness of the intervention (e.g., the cost per child enrolled) and cost-benefit studies (e.g., the sum costs and benefits of the program). It is important that this information be linked in some way to the primary outcome studies so that it can be retrieved.

4. Study Team

Anthony Petrosino, Ph.D., is Senior Research Associate at Learning Innovations at WestEd, and Associate Director of Research for the Regional Education Laboratory, Northeast and Islands. Anthony has worked on a number of projects during the past 20 years to identify, retrieve, appraise, analyze and report on separate but similar studies. For example, he was one of the founding members of the Campbell Collaboration, assisting in the development of its first trials register (C2-Spectr), co-authoring its pilot review (on “Scared Straight” and other juvenile awareness programs), and serving as Founding Coordinator for its Crime and Justice Group. A version of the “Scared Straight” review received the prestigious Pro Humanitate Literary Award from the Center for Child Welfare Policy of the North American Resource Center for Child Welfare. Although most of his training and experience has been in the justice area, he has more recently been working in education, and has co-authored government reports on the school dropout issue and the use of interim assessment in low-performing schools in Massachusetts.

Claire Morgan, M.A., is a Research Associate at Learning Innovations at WestEd. She brings rich experience and sensitivity to issues facing developing nations. Morgan has lived and worked in Mexico, Central America, and the South Pacific, and has considerable experience conducting research in international issues and among marginalized populations, including a study of the education and work experiences of Tongan immigrants, action research around non-formal education of Latino immigrants, and current work on English language learners, education policy issues in Puerto Rico, and the achievement of Hispanic immigrant students in the U.S. Virgin Islands. Morgan serves as the lead researcher for Puerto Rico and the U.S. Virgin Islands for the U.S. Department of Education-funded Regional Educational Laboratory (REL) Northeast & Islands. In this capacity, she consults with the Departments of Education of Puerto Rico and the U.S. Virgin Islands and provides research and technical assistance to address their education policy priorities. Morgan’s other research and evaluation work includes evaluation of National Science Foundation (NSF) university-school partnerships, evaluations of the federally-mandated Supplemental Educational Services (SES) program for the Massachusetts Department of Education and of the English to Speakers of Other Languages (ESOL) program for Prince Georges County, Maryland, and an evaluation of a language-minority community schools initiative in Quebec, Canada. Prior to joining WestEd, Morgan, who is fluent in Spanish, developed and directed a community-based adult education program for Nuestra Casa, a nonprofit organization serving Latino immigrants in Northern California. This initiative, a partnership with an underserved school district serving a large Hispanic student population, increased parent participation in schools and contributed to community development. In addition, she completed an internship at International Development Exchange (IDEX) in San Francisco, providing support to micro-ventures in Latin America. Morgan is currently

directing the expansion of a nonprofit Latina women's work cooperative that she co-founded in 2006. She received an MA in International Education Administration and Policy Analysis from Stanford University. She is particularly skilled in designing and conducting quantitative and qualitative research and evaluation, including instrument development and coding, literature reviews, interviews and observations, and analyzing demographic and achievement data.

Robert Boruch, Ph.D., is University Trustee Chair Professor of Education and Professor of Statistics (Wharton School) at the University of Pennsylvania. He has also served as faculty in the Fels Center for Government and at the Annenberg School Statistical Institutes. He is principal investigator for the What Works Clearinghouse (US Department of Education) and co-chairs the Steering Group of the international Campbell Collaboration. Dr. Boruch is a leading expert in experimental design research methods and related science policy on estimating effects of interventions. He has advised governments, private foundations, and research firms on randomized field trials in education, criminal justice, employment and training, and social welfare in the US and in other countries. His earliest contributions, during the 1970s, included service as advisor on the Cali Colombia randomized trials on cultural enrichment programs and the Nicaraguan trials on radio based mathematics education. Boruch chaired the National Academy of Sciences Committee on Evaluation of AIDS prevention programs in the 1980s and contributed to the WHO committee on the topic during the same period. He has authored numerous books and peer reviewed articles on related subjects. The most recent products include Evidence Matters: Randomized Trials in Education Research (2001), edited by Mosteller and Boruch (Brookings Institution Press), and a special edition of the *Annals of the American Academy of Political and Social Sciences* on place randomized trials in developed and developing countries (May 2005, volume 599) which covers health, crime and justice, welfare, housing, and education. . Boruch has been leader in institutes on generating better evidence for the US National Academy of Sciences and the Israel Academy of Sciences and Humanities, the Campbell Collaboration, workshops/seminars for the World Bank' IPDET, and in other venues. Boruch is an elected Fellow of the American Statistical Association, the Academy of Experimental Criminology, and the American Academy of Arts and Sciences, and is a Lifetime Associate member of the National Academy of Sciences (US). He has received awards for the work from the American Evaluation Association (Myrdal Award), American Educational Research Association, the Campbell Collaboration, and the Policy Studies Organization.

5. Timeframe

| Task to be Completed | Target Date for Completion |
|----------------------------------------------|----------------------------|
| Search for published and unpublished studies | December 2009 |
| Data extraction from research reports | February 2010 |
| Statistical analyses | March 2010 |
| Preparation of C2 draft final review | May 2010 |

6. Plans for updating the review

We plan to update this review in 36 months, in concert with C2 guidelines.

7. Acknowledgments

We would like to thank International Initiative for Impact Evaluation (3ie) of the Global Development Network (GDN) for their support of this project.

We thank the following persons for their helpful comments on this protocol (in alphabetical order): Campbell Education and Methods Group peer reviewers, WestEd colleagues Mary Cazabon, Sarah Guckenbug, Sue Henderson, Daniel Mello, and Eliza Spang, and 3ie officials Hugh Waddington and Howard White.

8. Statement concerning conflict of interest

We do not have any conflicts of interest regarding school enrollment policies. None of the authors has any financial or other personal interest in the results of this review.

9. References

Birdsall, N. (1999). Comment: A vicious cycle. In V. Tanzi, K. Chu, & S. Gupta (Eds.), *Economic policy and inquiry*. Washington, D.C.: International Monetary Fund.

Birdsall, N., Levine, R., & Ibrahim, A. (2005). *Toward universal primary education: Investments, incentives, and institutions*. London: United Nations Development Programme.

Bobonis, G., Miguel, E., & Sharma, C.P. (2006). Iron deficiency, anemia, and school participation. *Journal of Human Resource*, 41(4), 692-721.

Boruch, Robert F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Brembeck, Cole S. (1962). Education for national development. *Comparative Education Review* 5 (3), 223-231

Cueto, S. & Chinen, M. (2008). Educational impact of a school breakfast programme in rural Peru. *International Journal of Educational Development*, 28(2), 132-148.

Duflo, E. and M. Kremer (2005). Use of randomization in the evaluation of development effectiveness. Pgs 205-247 in George Keith Pitman, Osvaldo N. Feinstein, and Gregory K. Ingham (editors) *Evaluating Development Effectiveness*. New Brunswick, NJ: Transaction.

Filmer, D. (1999). The structure of social disparities in education: Gender and wealth. Policy Research Report on Gender and Development, Working Paper Series, No. 5. Washington, D.C.: The World Bank.

Filmer, D. (2004). If you build it, will they come? School availability and school enrollment in 21 poor countries. World Bank Policy Research Working Paper 3340. Washington, D.C.: The World Bank.

Filmer, D. & Schady, N. (2006). *Getting girls into school: Evidence from a scholarship program in Cambodia*. Washington, D.C.: The World Bank.

Glewwe, P., & Miguel, E. (2008). The impact of child health and nutrition on education in less developed countries. In T.P. Schultz & J. Strauss (Eds.), *Handbook of development economics, vol. 4* (pp. 3561-3606). Amsterdam: North-Holland.

Glewwe, P., Kremer, M., & Moulin, S. (forthcoming). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics*.

Glewwe, P. and M. Kremer (2006). Schools, teachers, and education outcomes in developing countries. In E. Hanushek and F. Welsh (eds.), *Handbook on the Economics of Education. Volume 2*. London: North-Holland.

Hanuam, E., & Buchmann, C. (2004). Global educational expansion and socio-economic development: an assessment of findings from the social sciences. *World Development*, 33(3), 1-22.

Herz, B., & Sperling, G. (2003). What works in girls' education: Evidence and policies from the developing world. Washington, D.C.: Council on Foreign Relations.

Levy, Dan Matt Sloan Leigh Linden and Harounan Kazianga (2009). *Impact Evaluation of Burkina Faso's BRIGHT Program* (MPR 6275-003). Princeton, NJ: Mathematica Policy Research.

Lockheed, Marlaine E. Stephen C. Vail and Bruce Fuller. How textbooks affect achievement in developing countries: Evidence from Thailand. *Educational Evaluation and Policy Analysis*. 1986; 8(4):379-392.

López-Calva, L. (2008). Cash transfers, conditions, and school enrollment in Ecuador. *Economía* 8 (2), 71-77.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), 159-217.

Oliver S., A.M. Bagnall, J. Thomas, J. Shepherd, A. Sowden, I. White, J. Dinnes, R. Rees, J. Colquitt, K. Oliver, and Z. Garrett, (2008). *RCTs for policy interventions?: A review of reviews and meta-regression*. Accessed online on May 7, 2009 at http://www.pcpoh.bham.ac.uk/publichealth/methodology/docs/publications/JH09_Oliver_2008_RCTs_for_policy_interventions.pdf.

Schultz, T.P. (2004). School subsidies for the poor: Evaluating the Mexican PROGRESA poverty program. *Journal of Development Economics*, 74(1), 199-250.

The World Bank (2004). World development report 2004: Making services work for poor people. Washington, D.C.: World Bank.

The World Bank (2008). Tembon, M., & Fort, L. (Eds). *Girls' education in the 21st Century: Gender equality, empowerment and economic growth*. Washington, D.C.: The World Bank.

UNESCO (United Nations Educational, Scientific, and Cultural Organization) (2007). Education for all global monitoring report: Education for all by 2015: Will we make it? Paris: UNESCO.

Appendix A. List of bibliographic databases to be searched

- Academic Search Premiere
- Ask.com
- Australian Education Index
- British Education Index
- British Library
- CBCA Education
- Chalk's e-Library Collection
- Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR)
- Campbell Collaboration Prospective Trials Register (C2-PROT)
- EBSCO Megafile
- Econlit
- Education Index
- Eldis
- ERIC
- Expanded Academic ASAP
- Full-Text of OVID Journals
- Google and Google Scholar
- Health Technology Assessment Register
- Housing and Urban Development (U.S.) "HUD USER"
- IDEAS-RESEARCH PAPERS IN ECONOMICS <http://ideas.repec.org/>
- Index to Current Urban Documents
- Inside Info Plus (British Public Library)
- International Bibliography of Social Science
- ISI Web of Science

- JOLIS (World Bank, International Monetary Fund, International Finance Corporation)
- Medline
- National Clearinghouse of Child Abuse and Neglect (NCCAN)
- PAIS Archive
- PAIS International
- Periodical Contents Index
- PolicyFile
- Proquest Dissertations and Theses
- Psychology and Behavioral Sciences Collection
- PsycInfo (includes PsychLit)
- Sage Family Studies Abstracts
- Sage Journals Management and Organizational Studies Full-Text Collection
- Sage Journals Political Science Full-Text Collection
- Sage Journals Sociology Full-Text Collection
- Sage Urban Studies Abstracts
- Social Service Abstracts
- Social Service Research Network
- Social Work Abstracts
- Sociological Abstracts
- UNESCO (UNESDO and UNESBIB)
- Wider Public Health Agenda Project Report (annotations of relevant reviews)
- Worldwide Political Science Abstracts
- Yahoo!

Appendix B. CODING INSTRUMENT

Coder:

- Claire Morgan
- Anthony Petrosino
- Bob Boruch
- Other _____

Citation for Primary Document:

I. RESEARCHER AND STUDY CHARACTERISTICS

What year was the primary document published?

How many documents were considered in coding this study?

What was the type of document?

- Book
- Book Chapter
- Government Report
- Journal (peer reviewed)
- Dissertation
- Unpublished (tech report, conference paper)

In what country did the evaluation take place? _____

World Bank country classification at time of study

- Lower Income
- Lower Middle Income
- Upper Middle Income

What was the setting for the evaluation?

Who conducted the evaluation? (e.g., medical researchers, economists, etc.)

What other information was provided on the context for the evaluation?

II. STUDY METHODS AND METHODOLOGICAL QUALITY

Was random assignment used to assign groups? (Yes/No)

At what level was randomization conducted?

How was the randomization specifically done?

Were there any randomization problems noted? (Yes/No)

If yes, please detail those problems below:

If random assignment was not used, what quasi-experimental method was used to equate groups? (e.g., matched comparison schools; post-hoc statistical matching of individuals; regression discontinuity; propensity scores; etc.)

Where did comparison group come from?

At what level was non-random assignment made?

Were any substantive differences in pretests of group equivalence noted?

(Yes/No)

If yes, please detail those differences below:

Were there any overall attrition problems noted? (Yes/No)

Was differential attrition noted? (Yes/No)

If yes, please detail those problems below (especially the magnitude of attrition, both from original sample and differentially between treatment and control groups):

How were attrition problems dealt with by investigators?

III. INTERVENTION AND CONTROL CONDITIONS

Number of groups in the study:

Rationale for selecting intervention and control contrast if multiple groups:

List excluded study groups with brief description:

Describe the intervention below, with particular attention to the 'dosage' of the treatment:

How many participants were randomized to this group?

**Were program implementation problems described by investigators?
(Yes/No)**

Detail fidelity problems below:

Please detail program theory (or mechanisms for why it should work):

What is the control or comparison condition?

- No Treatment Group
- Treatment as Usual Group
- Lesser but Innovative Treatment

Describe the control or comparison condition (including 'dosage' if applicable):

How many participants were randomized to this group?

IV. PARTICIPANTS IN THE STUDY

Type of school

Age/school level/grade

Percentage of participants that were female

Poverty/SES

Other data on participants (e.g. health, child labor, past enrollment status, achievement level)

V. OUTCOMES (SEE SHEET ON NEXT PAGE)

Include all data on treatment and control, including results, sample sizes used in analysis, the statistical technique, whether regression-adjusted or not, (and if so, what controls were used), statistical significance and probability level.

| Outcome | First Effect | Middle Effect | Last Effect |
|------------------------------|---------------------|----------------------|--------------------|
| Enrollment/ Matriculation | | | |
| Attendance | | | |
| Dropout | | | |
| Test Scores | | | |
| Grades | | | |
| Other (Indicate) | | | |

Please detail all subgroup effects below, particularly gender and grade level:

Please detail all cost/economic information below:

ANY OTHER COMMENTS ON THE PROGRAM OR EVALUATION