Identifying Effective Education Interventions in Sub-Saharan Africa:

A meta-analysis of rigorous impact evaluations

Katharine Conn

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Identifying Effective Education Interventions in Sub-Saharan Africa:

A meta-analysis of rigorous impact evaluations

Katharine Conn

The aim of this dissertation is to identify effective educational interventions in Sub-Saharan African with an impact on student learning. This is the first meta-analysis in the field of education conducted for Sub-Saharan Africa. This paper takes an in-depth look at twelve different types of education interventions or programs and attempts to not only present analytics on their relative effectiveness, but to also explore *why* certain interventions seem to be more effective than others. After a systematic literature review, I combine 56 articles (containing 66 separate experiments, 83 treatment arms, and 420 effect size estimates), and I use random-effects meta-analytic techniques to both a.) evaluate the relative impact of different types of interventions and b.) explain variation in effect sizes within and across intervention types.

When I examine the relative pooled effect sizes of all twelve intervention areas, I find that interventions in pedagogical methods (changes in instructional techniques) have a higher pooled effect size on achievement outcomes than all other eleven intervention types in the full sample (e.g., school management programs, school supplies interventions, or interventions that change the class size or composition). The pooled effect size associated with these pedagogical interventions is 0.918 standard deviations in the full sample (SE = 0.314, df = 15.1, p = 0.01), 0.566 in the sample excluding outliers and including only randomized controlled trials (SE = 0.194, df = 11, p = 0.01), and 0.228 in a sample that includes only the highest quality studies (SE = 0.078, df = 5.2, p = 0.03). These findings are robust to a number of moderating factors. Using meta-regression, I find that on average, interventions in pedagogical methods have an effect size

over 0.30 standard deviations (significant at the 5% level) greater than all other intervention areas combined, even after controlling for multiple study-level and intervention-level variables. Beyond this average effect, I show that studies that employ adaptive instruction and teacher coaching techniques are particularly effective. Further, while studies that provide health treatments or school meals have on average the lowest pooled effect size, I show that if these studies are analyzed using cognitive assessments (tests of memory and attention), health treatments actually produce a relatively large pooled effect size of 0.176 standard deviations (SE = 0.028, df = 2.18); this is particularly true of studies that either prevent or treat malaria.

In addition, this meta-analysis examines the state of current education impact evaluation research in Sub-Saharan Africa and highlights both research gaps as well as differences in study design, methodology, and reporting of metrics by academic field. I find that the bulk of the research in this area comes from the field of economics (62%), followed by the fields of education (23%) and public health (15%). Further, the majority of this research has been conducted in a set of six countries: Kenya, Nigeria, South Africa, Uganda, Burkina Faso, and Madagascar, while rigorous evaluations of education programs have never taken place in others. Moreover, topics currently under rigorous study are not necessarily representative of the major issues facing many Sub-Saharan African school systems today. For example, there are no impact evaluations of multi-grade or multi-shift teaching and only one evaluation of a bilingual education program. This meta-analysis thus recommends a shift in the impact evaluation research agenda to include both a broader geographic and topical focus, as well as an increased emphasis on improvements in pedagogical methods, without which other interventions may not reach their maximum potential impact.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge several individuals that have made this research possible. Most especially, I would like to thank my sponsor, Elizabeth Tipton, for her constant encouragement, patience, and expertise. I would also like to thank my other committee members Thomas Bailey (Chair), Judith Scott-Clayton, Patrick McEwan, and Douglas Ready, as well as Henry Levin and Francisco Rivera-Batiz for their thoughtful suggestions and comments. I am grateful also to Harry Patrinos, Emilio Porta, Oni Lusk-Stover, Jian Guo Zhu, Husein Abdulhamid and the World Bank's HDNED team who made the data collection behind this meta-analysis possible. Thank you also to SREE Spring 2014 conference attendees, the Economics and Education seminar attendees (Teachers College, Columbia University), and to Lauren Fellers for helpful feedback and assistance. Finally, I thank my parents, family and friends for supporting me and Brendan Manquin for believing in me.

DEDICATION

For my students

~ Timbi Touni, Guinea

CHAPTER I.

INTRODUCTION

The objective of this research is to highlight which education interventions are most effective at improving learning outcomes across Sub-Saharan Africa and why – a critical policy imperative for developing countries with limited resources to spend. For example, do financial incentives for teachers result in similar student achievement outcomes as providing incentives to students or their families? Or what explains the variation in the impact of school-based management interventions? To answer such questions, this paper takes advantage of a new wave of research in the social sciences: rigorous "impact evaluations" that employ experimental or quasi-experimental methods to measure the causal effect of programs through the creation of comparable treatment and control groups.

Previous reviews have attempted to synthesize this growing body of impact evaluations in developing countries, though not specifically for Sub-Saharan Africa (Kremer and Holla, 2009; Glewwe, 2011; Kremer, Brannen, and Glennerster, 2013; McEwan, 2013, and Krishnaratne, White and Carpenter, 2013). These reviews contain between 9 and 19 studies that take place in Sub-Saharan Africa; this dissertation expands significantly on this work by synthesizing over 56 education impact evaluations (containing 66 separate experiments and 83 treatments) which have been conducted in recent years in Sub-Saharan Africa. These impact evaluations are largely randomized controlled trials (43 studies containing 65 experiments), but quasi-experimental studies are also included (using propensity score matching, regression discontinuity, difference-in-differences, and instrumental variables designs). In general, this sample of studies is relatively large, given that 50% of recent meta-analysis in education have contained less than 40 studies (Ahn, Ames, and Myers, 2012, as cited in Tipton (*in press*)).

Further, the set of studies under analysis in this meta-analysis covers over twelve different intervention types, ranging from pedagogical interventions, to interventions in school health, to programs that alter student or teacher incentives. Table 1 (below) includes a full list and explanation of program types under study. These studies are grouped by channels that could be considered binding constraints to learning. For example, if school or system accountability is poor, programs that increase information, transparency, and monitoring may produce large impacts. Alternatively, if student or teacher motivation is low, programs that provide incentives for attending school and attaining high performance levels may have high returns. A full account of each of these channels through which learning levels could be altered is provided in Chapter III ("Theory of Change"). Also note that the systematic search described in Chapter IV covers additional topics, but experimental and quasi-experimental studies were found only in the areas listed above.

This dissertation combines robust meta-analysis techniques with supplemental narrative reviews as a synthesis method. Meta-analyses have been used in education research within the United States and Europe for decades (e.g., Glass and Smith, 1979; Hedges, Laine, and Greenwald, 1994; Means et al, 2010) but much less often in the developing world context. In fact only three meta-analyses (with only two on focusing on learning outcomes) have been conducted in the developing world in total (McEwan, 2013; Krishnaratne, White and Carpenter, 2013 and Petrosino et al, 2012). This paper contributes to this literature as it is the first meta-analysis in education for Sub-Saharan Africa.

Table 1. Intervention areas included in pan-African meta-analysis

Intervention Area:	Explanation:
Quality of Instruction	
Class Size & Composition	Class size reduction/ increase; Grouping of students by ability
Instructional Time	Increase/decrease in the number of instructional hours
Pedagogical Interventions	Programs that affect the method(s) of instruction & learning (includes blended or technology-assisted learning)
School Supplies Provision	Provision of flipcharts, textbooks, writing materials etc.
Student or Community Financial Limitations	
Abolishment of School Fees	Elimination of school fees
Cash Transfers	Programs which offer monetary support to families (can be conditional on their children's school enrollment)
School Infrastructure	Programs that construct/renovate schools
School or System Accountability	
Information for Accountability	Interventions that provide student/ school performance information to parents/ communities to increase transparency
School-Based Management/ Decentralization	Devolution of certain powers/responsibilities regarding education provision to the school-level (parents or school community etc.)
Student Cognitive Processing Ability	
School Meals	Programs which include nutritional supplements/school meals
Health Treatments	Programs including treatment for malaria or helminth infections (de-worming)
Student or Teacher Motivation	
Student Incentives	The provision of scholarships (or other incentives) to students
Teacher Incentives	Teacher performance-based pay; Temporary teacher contracts

Meta-analysis methods allow us to systematically review literature by standardizing and statistically combining the findings of similar evaluations, thus increasing statistical power and allowing us to estimate an average effect size. Meta-analytic methods will be used in this dissertation to 1.) explore the current state of the literature in this field, 2.) examine the relative effectiveness of different types of education interventions, and 3.) explain heterogeneity in effect size measures both across and within intervention areas.

This meta-analysis proceeds as follows: the remaining portion of Chapter II provides the motivation for this research, along with a review of relevant research. Chapter III provides an overview of the research objectives and theory of change. Chapter IV details the methods used to conduct a systematic search of the literature, calculate comparable effect sizes across studies, and estimate a random-effects meta-analytic model. Chapter V presents the results of this meta-analysis in three main parts: Part A evaluates the state of the literature (amount of research conducted by topic and by region, along with a summary of descriptive differences by academic field). Part B explores the relative effectiveness of each intervention area, identifies intervention types that appear particularly effective or ineffective, and tests the robustness of these findings. Part C examines the extent to which variation between effect size measures can be explained by characteristics of the intervention itself or by additional contextual factors. The last part of Chapter V includes a discussion of publication bias within this literature (Part D). Finally, Chapter VI concludes with an overview of the findings of this meta-analysis, a comparison of these findings to those of other authors, a note on the limitations of this analysis, and implications for future practice and research.

CHAPTER II.

MOTIVATION & RELEVANT LITERATURE

This chapter presents the motivation for this research agenda in the Sub-Saharan African context. Both the policy relevance of this work and the choice of meta-analytic methods to best assess the relative effectiveness of this experimental and quasi-experimental literature are discussed. Further, this chapter provides an overview of the findings of other research syntheses that have attempted to aggregate some of this evidence world-wide, but not for Sub-Saharan Africa specifically. Finally, the contribution of this meta-analysis to the literature is discussed.

A. Context and Motivation for this Research

Current learning levels for primary as well as secondary school students are extremely low in much of Sub-Saharan Africa. Among the few Sub-Saharan African countries that participate in international assessments, performance rates have been consistently in the bottom 5-6% of countries for both fourth and eighth grade students (TIMSS 2003, 2007, and 2011).[1] Further, internationally administered early-grade diagnostic tests of reading and math such as the Early Grade Reading Assessment (EGRA) and the Early Grade Math Assessment (EGMA), show extremely low levels of proficiency among lower primary school students. For example, an EGRA diagnostic in three provinces in South Africa found that 65.2% of grade 1 students sampled were not able to identify a single letter sound at baseline (Piper, 2009); EGMA results from a diagnostic test in Bauchi, Nigeria indicate that only 20% of second grade students are able to correctly identify numbers (USAID, 2013), and a 2011 EGRA diagnostic of students in

[1] Institute of Education Sciences, National Center for Education Statistics, Trends in International Mathematics and Science Study (TIMSS). http://nces.ed.gov/.

grade 2 and 3 in Liberia found baseline levels of oral reading fluency to be only 20 words per minute (Piper, 2011).

Meanwhile, progress over the last decade in regards to school access and enrollment has been promising. Between 1999 and 2009, the primary gross enrollment rate grew by 3.1 percentage points per year (Majgaard and Mingat, 2012), and by 2011 the net enrollment ratios for lower and secondary and primary schools were 82% and 89%, respectively (UIS-UNESCO, 2014). Numerous government agencies, NGOs, international organizations, and campaigns have supported this focus on access, including the United Nations (Millennium Development Goals) and the Education for All Initiative (The World Bank, UNESCO, development agencies, national governments, and civil society groups). However, low student attendance and retention rates continue to be an issue. In 2009, the average primary completion rate across the Sub-Saharan African region was 67% (Majgaard and Mingat, 2012). Majgaard and Mingat (2012) suggest that these drop-out rates are indicative of education systems that are not providing the right type or quality of education that students and parents demand. Given these quality issues and the low levels of performance noted above, there has also been a renewed focus on learning outcomes in Sub-Saharan Africa, both within countries as well as from international organizations (Gove and Cvelich, 2011 and UIS-UNESCO, 2014). And in fact, increasing learning outcomes may contribute more to economic growth than school enrollment alone (Hanushek and Wöessmann, 2007).

An experimental literature has developed over the last three decades which attempts to address this need for improved learning outcomes in Africa (as well as in other developing countries). These "impact evaluations," which are predominately randomized controlled trials (though can also refer to quasi-experimental designs) have taken place all over the globe, including a large number in Sub-Saharan Africa. For example, Blimpo and Evans (2011) examine the impact of school-based management training on subsequent student achievement in The Gambia; Piper and Korda (2011) estimate the impact of adaptive instruction & teacher coaching on student learning in Liberia; and Kremer, Miguel and Thornton (2009) examine the impact of merit scholarships on student achievement in Kenya.

Because these experiment-based evaluations have the potential to offer increased internal validity, it has been argued that "randomized studies offer the most promise for understanding the impact of policies on learning" (pg. 946) (Glewwe & Kremer, 2006). However, these experimental studies have been criticized for two main reasons which relate to their generalizability: 1.) these experiments are too narrow in focus and too specific to a particular locale in order to generalize their results and 2.) these experiments often lack the ability to isolate the true causal mechanisms behind a program's impact (Deaton, 2010; Ludwig, Kling, and Mullainathan, 2011; and Angrist and Pischke, 2009, 2010).

In order to address the generalizability of these results, this dissertation uses meta-analysis techniques to synthesize this new evidence. Meta-analysis methods allow us to systematically review a literature by standardizing and statistically combining the findings of similar evaluations in a way that is transparent and replicable. Even when studies differ by study populations, intervention characteristics, or methods, random-effects meta-analysis can be used to a.) estimate the average effects of different types of interventions across contexts, b.) test the robustness of these results (controlling for various moderator variables such as study locale, length of intervention, or study methodology), c.) quantify the variation that exists across studies (if there is little, then it may be that certain interventions seem to be successful regardless of contexts etc.), and d.) help explain what contributes to this variation (for example differences may be due to the scale of a program or differing initial conditions of a system or population). Explaining this variation in findings can also lead to the isolation of causal mechanisms if certain aspects of a treatment differ between two otherwise similar programs. Importantly, meta-analysis also sheds light on the state of the literature in a particular field and highlights research gaps.

B. Findings of other syntheses

Reviews of this new body of experimental research are still quite limited and none focus on Sub-Saharan Africa specifically; this dissertation is in fact the first meta-analysis to do so. A number of books by World Bank researchers have provided narrative syntheses of this research by topic, including the literature on conditional cash transfers (Fiszbein & Schady, 2009), teacher incentives (Vegas, 2005), school-based management (Barrera-Osorio et al, 2009), and interventions that improve accountability (Bruns, Filmer, and Patrinos, 2011). Further, reviews which explore the impact of multiple types of programs on student learning outcomes in the developing world overall include: Kremer and Holla (2009) who conduct a narrative review of approximately 65 randomized controlled trials; Glewwe, Hanushek, Humpage, and Ravina (2011), who review 79 correlational, experimental, and non-experimental studies; Kremer, Brannen, and Glennerster (2013) who review a selection of 18 randomized controlled trials; McEwan (2013) who conducts a meta-analysis of 76 randomized experiments; and Krishnaratne, White and Carpenter (2013) who conduct a meta-analysis of 75 experimental and quasi-experimental studies. These reviews contain between 9 and 19 studies conducted in Sub-Saharan Africa. Below I lay out the basic conclusions of authors who have conducted either narrative, vote-counting, or meta-analytic reviews of the evidence.

Narrative reviews: Kremer and Holla (2009) review a selection of approximately 65 randomized controlled trials in the developing world (containing approximately 12 education impact evaluations taking place in Sub-Saharan Africa). They find that technology-assisted learning, remedial education, student tracking and the use of contract teachers were among the most effective interventions at improving student achievement. Kremer, Brannen, and Glennerster (2013) review a selection of 18 randomized controlled trials (containing 9 impact evaluations taking place in Sub-Saharan Africa).[2] Their findings regarding student learning are similar to Kremer and Holla (2009), with the additional note that technology-assisted learning and remedial education are particularly effective when the intervention is adaptive to the student's learning level. They also find that inputs such as hiring additional teachers, buying more textbooks, or providing flexible school grants have surprisingly small effects.

Vote-counting review: Glewwe et al (2011) conduct a "vote counting" synthesis (authors sum statistically significant/ insignificant positive and negative findings) of studies with educational outcomes in developing countries from 1990-2010 (with 17 studies taking place in Africa), including both correlational and experimental studies. They find variables associated most frequently with statistically significant effects include infrastructure-related studies (quality of roof, availability of desks & chairs etc.), having a longer school day, teacher knowledge of the subjects they teach, and teacher absence. However, they find that most of the effect sizes in their analysis are statistically insignificant, which limits their interpretation of these interventions. Furthermore, as Hedges, Laine, and Greenwald (1994) have shown in a U.S. based analysis, there are methodological limitations to vote-counting as a statistical procedure, particularly in terms of statistical power.

Meta-Analyses: McEwan (2013) conducts a meta-analysis of 76 randomized controlled trials from the developing world, 19 of which take place in Africa. He finds that intervention

[2] This analysis also examines the cost-effectiveness of various interventions.

types with the largest average effect sizes are those that employ computer-assisted learning, teacher training, and smaller classes/smaller learning groups/ ability grouping (with effect sizes ranging from 0.12 to 0.15 standard deviations). Intervention types that are among the least effective included school monetary grants and nutritional/ health treatments (effect sizes between 0.04-0.06 standard deviations). He also finds evidence which suggests that the effect sizes associated with school materials and contract teachers are compounded by effects of co-occurring teacher training and class size reduction programs.

A second meta-analysis was conducted by Krishnaratne, White & Carpenter (2013) which includes 75 studies (both experimental and quasi-experimental), 9 of which take place in Africa. They find that school supplies materials had the highest known impact (but only for Math scores, not Language), and they state that school-based management, teacher resources & school meals are all promising interventions. They state that they don't have enough data to report on the relative impact of the remaining interventions. However, while this review provides a relatively detailed narrative of the variation in program effects within the body of the paper, the statements on "effective" or "ineffective" interventions in the conclusions section are based on relatively little data; they state that "when over six or more studies pooled together (have) effect sizes that (are) positive and significant at the 10% level," this group of interventions (is) labeled "what works," while "when more than six studies pooled together (show) no significant impact," this (is) labeled as "what doesn't work." (Figure 2, pg. 42). Because these findings are based on little data and because authors do not use robust variance estimations or formally examine program heterogeneity, these prescriptions may be premature.

C. Contribution of this meta-analysis

This dissertation is the first meta-analysis in education to focus on Sub-Saharan Africa and only the third to focus on learning outcomes in developing countries overall. This is in contrast to the over 200 meta-analysis conducted in education in other continents and countries world-wide (the majority of which are in United States alone).[3] The fact that the developing world is socially and economically diverse and that different interventions and programs are often culturally specific suggests that the broad results from the three previous syntheses may not be applicable to Sub-Saharan Africa, which is itself a quite diverse region. Thus the extent to which the findings of other syntheses are consistent with my results in Sub-Saharan Africa is a question I examine in a later section of the paper (Chapter VI). As mentioned above, current syntheses of this literature contain between 9 and 19 studies, while this meta-analysis contains 55 studies (66 separate experiments), as I extended the reach of this meta-analysis to include studies from the fields of economics, public health, as well as education (all with an impact on student learning). I also include quasi-experiments in my sample. Finally, I use new meta-analysis estimation techniques such as robust variance estimation (RVE) and small sample corrections (see Chapter IV) to ensure the accuracy and interpretability of my findings.

[3] Results from ProQuest search of ERIC database on 04/05/2014. Search included journal articles, books, reports, and dissertations for meta-analyses in the field of education (meta-analysis in title).

CHAPTER III.

RESEARCH OBJECTIVES & THEORY OF CHANGE

The three major research questions examined in this dissertation relate to 1.) the state of the literature in this field, 2.) the relative impacts of different types of interventions, and 3.) how variability in these impacts can be explained. Below I provide an overview of each question. Further, I present below a "Theory of Change" as it relates to this research; in this section I explain how different types of interventions included in this analysis are expected to influence student learning. This "Theory of Change" identifies five actionable areas through which learning could be improved: quality of instruction, student cognitive processing abilities, student or teacher level motivation, school or system level accountability, and student or community financial limitations.

A. Research Objectives

The overarching question examined in this paper is the following: which interventions have the highest impact on student learning outcomes in Sub-Saharan Africa? To do this, I examine 56 impact evaluations (containing 66 separate experiments and 83 treatments) covering numerous interventions areas (ranging from class size reduction programs - to conditional cash transfer programs - to interventions in school health/ nutrition), and I employ meta-analysis techniques which allow to me aggregate and analyze this data in a way that is transparent, structured and provides for study-to-study comparability. This paper is then organized around a set of three main research questions (below).

1.) What is the state of the literature in this area? Though a systematic search process, as prescribed by meta-analytic methods and described below in Chapter IV, I am able to evaluate gaps in the literature from both a regional and topical perspective. In addition, I ask questions of this body of literature overall such as, what are the main methodologies used to study these effects, are these studies nationally representative, and do these studies evaluate the psychometrics of their assessments used to measure student performance? Finally, I examine whether different academic fields that have contributed to this literature (economics, education, and public health) differ in terms of methodological design, reporting of study findings, region under study, assessment type, publication type (journal versus working paper or report), and the size of the experiment (among other variables).

2.) What is the relative effectiveness of different intervention types? Secondly, I evaluate the average impacts of different types of education interventions on student learning (and associated levels of heterogeneity) and identify interventions that appear to be particularly effective or ineffective. I then test whether these findings are robust to the inclusion of study moderators such as region, quality, assessment type, and methodology.

3.) What explains variation in effectiveness? Finally, I attempt to explore variation in effect size differences both across and within intervention types. For example, I ask whether studies that employ standardized assessments have larger or smaller pooled effect sizes on average than researcher-created tests or whether studies that employ matching methodologies have a statistically significantly different pooled effect size than those using RCTs or other quasi-experiments. Further, I explore heterogeneity within each intervention type, identifying which characteristics of an intervention or treatment mechanisms may have contributed to differential study impacts.

B. Theory of Change: Channels that influence learning

The studies included in this meta-analysis all affect student learning though one or more of the following five channels: quality of instruction, student cognitive ability, student or teacher motivation, school or system-level accountability, and student or community financial limitations. Here, I explore the types of studies found in each category and the ways in which these interventions could have an impact on student learning. Again, the extent to which any of these channels has an impact on student learning may depend on an area's initial conditions or specific binding constraints.

Quality of instruction: Interventions that alter the quality of instruction include the provision of school supplies such as textbooks or other instructional materials (flipcharts, notebooks etc.), changes in class size or the use of ability grouping (tracking), increases in instructional time, and changes to teachers' pedagogical techniques. Textbooks and other school supplies should in theory increase student comprehension of course content and enable additional practice in math and language lessons, assuming that the textbooks are at the appropriate level for students, that they contain curricula-relevant information, and that teachers integrate the use of these textbooks into their lesson plans. Class size interventions, as well as student ability-tracking can allow teachers to better tailor their instruction to individual students, smaller groups of students, or students of similar learning levels, enhancing student performance; this again assumes that teachers know how to fully take advantage of smaller class sizes and ability-tracked groups in ways that can maximize student learning. Teachers may also prefer teaching to smaller and more homogenous groups of students, which may also affect teacher attendance, and thus student learning. Finally, employing pedagogical techniques that are more engaging and driven by education theory such as visual learning, cooperative learning, or procedural learning

(learning which emphasizes the mastering of skill sequences), may (if teachers are well trained), influence student performance. This is particularly true in schools were students are accustomed to lecture-based rote learning.

Student cognitive processing ability: For students with helminth infections (worms), suffering from malaria, or suffering from malnutrition or under-nourishment, programs that address these issues are thought to have a direct effect on student performance through increased cognitive processing abilities. These treatments include anti-helminth or anti-malarial treatments, as well the provision of school meals (breakfast, lunch, or mid-morning snacks) and nutritional supplements. School meal programs in particular, may also influence student attendance and could thus improve student learning through this vein as well.

Student or teacher motivation: This category includes studies that focus on student or teacher performance incentives. Student incentives such as merit scholarships (Kremer, Miguel, and Thornton, 2009) or monetary prizes for achieving a performance target (Blimpo, 2010) are theorized to affect performance through increased student effort; it may also be that these incentives increase the likelihood that students attend school more frequently, which could again contribute to higher learning levels. Teacher incentives are likewise thought to increase student performance due to increased teacher effort either in or outside of the classroom (more time spent with individual students or more care taken to prepare or present lessons), or due to the fact that teachers under such incentives are more likely to attend classes and thus increase student learning through increased instructional time. In addition, the offering of certain incentives may bring new teachers to the profession and change the make-up of education professionals; this could result in higher student performance if these new teachers are able to teach effectively.

School or system-level accountability: Interventions that are theorized to work through increased accountability include management interventions (at the school, sub-district, or district level), as well as information interventions targeted at the school community and households. These interventions are expected to influence student performance through a variety of channels. First, programs that empower local school committees (allowing them to monitor/ hire/ fire teachers or administer resources), may put pressure on teachers to attend school regularly and improve student performance. These community-based interventions may also work through increased parent participation in the student's education (increased supervision of homework completion or increased enforcement of student's school attendance). Informational and management tools provided to schools and districts (including school and district report cards) may also help schools to be run more efficiently and encourage school communities to track student progress, which may also put pressure on both principals and teachers to improve performance results. Information interventions that make transparent the amount of public funding that is supposed to arrive at a school may also increase student performance if communities keep schools accountable for the use of funds (textbooks and supplies to students). And community information programs that emphasize to students the real rate of return to education may help both students and households recognize the importance of student retention and performance for future payoffs.

Student or community financial limitations: Finally, in some communities, students may be motivated to attend school but lack the financial means to gain an education, either due to the fact that they are not able to pay their fees or because they do not own a uniform, a requisite for the school-going population in some countries. Programs that address such issues include cash transfers programs that pay for students to attend school, national policies which abolish school fees, or programs that provide uniforms to students. These policies and programs all have the potential to increase not only enrollment and attendance, but also to decrease drop-out rates. These programs may also free up household income to put towards educational supplies if tuition itself is no longer a financial constraint, all of which would presumably increase student achievement.

CHAPTER IV.

RESEARCH METHODS

Meta-analytic methods follow a structured protocol regarding the systematic search for literature, the coding process, effect size calculations, and the choice of estimation method. The following chapter first provides an overview of meta-analysis, as well as its limitations (Section A), then explains in detail: how studies were identified for inclusion in this dissertation (Section B), how a thorough literature search was conducted (Section C), how information was extracted and coded from each of the primary studies (Section D), and finally, which estimation models were used to aggregate and analyze this data (Section E).

A. Overview of Meta-Analysis

Meta-analysis provides a way of statistically aggregating study results that is transparent, replicable, and allows for study-to-study comparability through the standardization of findings. Because meta-analysis enables us to statistically combine findings of similar evaluations, this increases the statistical power of the aggregate measure, allows us to empirically estimate overall effect sizes, and to better explore effect size variation between studies. Meta-analysis provides researchers with not only pooled effect size estimates, but also ways to check the robustness of these estimates and explore to what extent variability in these estimates could result from observable moderating factors. Further, through its structured approach, meta-analysis enables us to find effects or see patterns that we would not have seen using other synthesis methods, and it forces us to realize how much we know or don’t know about an area of study.

Other synthesis methods such as narrative reviews and “vote counting” methods have been criticized by researchers for a variety of reasons. Narrative reviews are said to be overly subjective and biased in favor of the research or methodology types with which the author is most familiar or most interested (Light and Pillemer, 1982). In addition, “vote-counting” syntheses which sum the number of positive significant, positive insignificant, negative significant, and negative insignificant results have been criticized as ignoring both the magnitude of the findings and discriminating against small experiments, which are most likely to have large standard errors (Cook and Leviton, 1980; Glass, McGaw, and Smith, 1981; Jackson, 1980; and Light and Smith, 1971).

Meta-analytic methods can be used to overcome these issues both through the aggregation of results across studies (meta-analysis procedures allow small studies to be pooled together, thereby increasing statistical power) and also due to the reliance of meta-analyses on a set of transparent systematic search, coding, and reporting standards. These standards are described and promoted by The Campbell Collaboration[4] (for the social sciences) and The Cochrane Collaboration[5] (for the natural sciences, particularly medicine), which also house on-line libraries of peer-reviewed meta-analyses that have met these standards. For example, these standards require that coding protocols are developed, that multiple coders are used, and that inter-rater reliability is measured (and is high) in order to ensure that a researcher’s bias has not influenced study coding.

Meta-analyses typically report not only overall pooled effect sizes (for both the full sample and sub-groups) but also heterogeneity statistics. These include an estimate of τ^2 which is a measure of true heterogeneity between studies (not due to chance) and I^2, which measures the percentage of the total heterogeneity due to true variation. In addition, meta-analyses report

[4] http://www.campbellcollaboration.org/

[5] http://us.cochrane.org/

metrics which indicate the presence (or absence) of publication bias. These standards include the use of funnel plots and the Egger test (Egger, Smith, Schneider and Minder, 1997). Funnel plots provide a visual diagnostic of publication bias by showing gaps in the number of small studies with small effects (or effects to the left of the average pooled effect with high standard errors), which are less likely to be published. The Egger test is a formal test of publication bias; it detects asymmetry in the funnel plot by examining a regression of the standardized effect estimates against their precision and determining whether the intercept deviates significantly from zero.

The use of meta-analysis is not without its limitations, although a number of these limitations do pertain to any synthesis techniques available. For example, though studies found in meta-analyses are identified through a comprehensive systematic search of the literature, it is still the case that not all programs that have been implemented in this field have been studied. Further, those that have been studied do not necessarily represent a random sample of these programs overall. In addition, meta-analysis suffers from some practical limitations that are not inherent to meta-analysis itself but rather to the state of the literature that the meta-analysis is describing. First, while meta-analysis itself provides a way to test the external validity of a group of findings, when samples are small, the ability to technically account for moderating factors may decrease. Thus if studies are few and their characteristics are relatively heterogeneous, providing a supporting narrative review is necessary in order to account for study-to-study variation (Waddington, 2012). In addition, studies may not supply all of the necessary moderator variables to be included in an analysis, and some of these factors may not be known (e.g., certain initial conditions or the binding constraint of a region’s education system). Further, even if the number of studies is quite large (e.g., over 200), and all studies consistently supply the same moderator variables, the adjusted analysis is still correlational in nature, even if the studies within the meta-analytic sample are all causally estimated.

Importantly, while meta-analysis addresses some features of generalization, it does not address all features, particularly those related to partial or general equilibrium issues. This is because RCTs in particular are often criticized for being at too small a scale to allow for generalizability (offering only a partial equilibrium result) (Kremer and Holla, 2009), and while the scale of a study is a moderator variable that could be included in the meta-analysis, it is unlikely that a large enough number of RCTs will be measured in both a partial and general equilibrium setting to control for this difference. However, I argue that a carefully constructed synthesis of partial equilibrium results is still a useful tool, as a number of these experiments may be local in their future implementation as well. Further, some intervention types may be more susceptible to this critique such as studies of the “private school advantage,” which I collect during my systematic search but do not analyze in my main sample due to the fact this “advantage” may say more about the current quality of public schools in general equilibrium than about the effectiveness of private institutions; in addition these studies do not represent an intervention or an actionable policy change, as do the others in my sample.

Finally, meta-analysis methods require a focus on standardization, systematization, and transparency at all stages of the data collection and analysis. In the next sections, I explain in detail how I identified studies for inclusion in this dissertation (Section B), how I conducted a thorough literature search (Section C), how I extracted and coded information from each of the primary studies (Section D), and the estimation models used to aggregate and analyze this data (Section E).

B. Identification of study selection criteria

Inclusion criteria: The inclusion criteria for study selection was set by limiting my search to studies in Sub-Saharan Africa from 1980 to present, focusing on students in formal education, and using rigorous experimental or quasi-experimental methods; these included randomized controlled trials, difference-in-difference specifications, instrumental variables methods, matching methods (propensity score, non-parametric or simple covariate matching), regression discontinuity designs or time series models w/ fixed effects (see Table 2 below). Studies contained in this analysis include those found in peer-reviewed journals, as well as in academic working papers or evaluation reports; including this "grey" literature can help to limit publication bias (full details of this search are described below). A wide search of social science, business, education, and health databases were conducted.

Table 2. Summary of Inclusion Criteria

REGION	Sub-Saharan Africa
PUBLICATION DATE	1980-present*
INTERVENTION AREA	See Table 3 (below)
OUTCOMES	One (or more) of the following outcome measures: learning outcomes**, repetition, drop-out, completion, retention, enrollment & attendance
METHODOLOGY	Experimental or quasi-experimental methods only: RCTs (randomized controlled trials), DID (difference-in-difference), IV (instrumental variables), Matching (propensity score, non-parametric or simple covariate matching), RD (regression discontinuity) or time series w/ fixed effects
PUBLICATION TYPE	Peer-reviewed journals, academic working papers or reports published through academic institution or research organizations
STUDY POP.	Any formal education level
FIELDS	Economics, Education & Public Health
TREATMENT COMPARISON	Comparing treatment to control/ "status quo"

*While I searched for studies from 1980 and onwards, 99% of studies in this sample were published after 2000.
**While the search was conducted for all outcomes, only learning/ testing outcomes were used in this meta-analysis.

Table 3 (below) details the intervention areas covered in this meta-analysis; they range from school meals programs to community information campaigns. The intervention areas are grouped by five channels through which learning outcomes may be altered: quality of

Table 3. Intervention Areas under systematic search:

Intervention Area:*	Impact Evaluation Examines:
Quality of Instruction	
After-School Tutoring	Additional instruction, targeting underperforming students
Class Size	Class size reduction/ increase
Technology-assisted learning	ICT computer programs employed both in the classroom as well as out-of-school
Instructional Time	Increase/decrease in the amount of instructional time.
Pedagogical Interventions/ Teacher Training	Programs that affect the method(s) of instruction & learning.
Language of Instruction	Changes in language of instruction policies locally/nationally.
School Supplies Provision	Provision of flipcharts, textbooks, writing materials etc.
Tracking & Peer Effects	Grouping of students by ability level; influence of certain students’ ability levels on rest of student population.
Student or Community Financial Limitations	
Abolishment of School Fees	Elimination of school fees
Cash Transfers	Programs which offer monetary support to families (can be conditional on their children’s school enrollment)
School Infrastructure	Programs that construct/renovate schools.
School Choice	Programs that offer households the right to choose their school
School or System Accountability	
High Stakes Testing & Accountability Systems	Advent of high-stakes testing regimes (such as NCLB in the US); advent of school/state accountability systems
Information for Accountability	Interventions that provide student/ school performance information to parents/ communities to increase transparency
School-Based Management/ Decentralization	Devolution of certain powers/responsibilities regarding education provision to the school-level (parents or school community etc.)
School Type	Impact of private, religious, single-sex, boarding, day schools.
Student Cognitive Processing Abilities	
School Meals	Programs which include nutritional supplements/school meals.
Health Treatments	Treatments for diseases such as malaria or helminth infections.
Student or Teacher Motivation	
Student Incentives	The provision of scholarships (or other incentives) to students
Teacher Incentives	Teacher performance-based pay; Temporary teacher contracts.

*While the search was conducted for all topics listed, the evidence was not found in a number of areas.

instruction, student or community financial limitations, school or system accountability, student cognitive processing abilities, and student or teacher motivation. For a full description of the ways in which these interventions could influence learning, see Chapter III, Section B. “Theory of Change” (above). Note that while the search was conducted for each of the areas listed in Table 3, studies in each category were not necessarily available (results of this search are detailed in Chapter V, Part A).

I limited my search to those studies with at least one education outcome. While I do include school-based health programs generally, I do not include school-based health studies which have *only* health outcomes. Also note that while my final analysis sample contains only those studies with learning outcomes, my search criteria includes all studies with any education outcome, given that some studies which seem to only examine enrollment or attendance, for example (as judged by their abstract), often include learning outcomes as well. Finally, I include studies with at least one treatment-to-control-group comparison.

Examples of studies included and excluded: An example of an “included” study is that of Christel Veermersch and Michael Kremer (2004) who conducted a randomized evaluation of a school meals programs in Kenya. The students involved in the study are enrolled in formal schooling (primary level), and the study is available as a World Bank Policy Research Working Paper. An example of study that was excluded is Garlick (2013), a tracking study, as it is the only tertiary education study in my sample and also does not take place in the classroom (the students were tracked by living quarters/ the dormitory). A second example of study that would be excluded is a journal article by Carnoy and Arends (2012) who explain mathematics achievement gains in Botswana and South Africa through associational methods; the authors examine correlations between classroom factors and student learning in grade 6 but do not establish an experimental or quasi-experimental identification strategy that would allow for causal claims.

Additional exclusion criteria: I exclude studies which identify themselves as randomized trials but randomize treatment between only two classrooms or schools, even if a student-level randomization between groups is conducted (examples include: Adegoke, 2011; Adeleke, 2007; Agbatogun, 2012; Akinsola & Animasahum, 2007; Awonyi & Ala, 1995; Onu, 2012, Plomp et al, 1991; Sarfo & Elen, 2007, and Talabi, 1989). In this case, student background characteristics across groups may be well-balanced on observables, but the impact of the intervention is still indistinguishable from the impact of the teacher or classroom, for example. For this reason, these methodologically unidentified studies are excluded. In addition, I also exclude studies that were not linked to an intervention, policy change, or treatment. For example, I exclude studies that measure the impact of malaria or the presence of intestinal worms on student performance but include studies that estimate the impact of malaria treatment or de-worming drugs on student performance (examples include Brooker et al, 2013 and Grigorenko et al, 2006).

C. Systematic literature search

An attempt was made to identify all studies pertaining to the research questions described above (within the corresponding inclusion criteria). Because impact evaluation research in the field of development economics is relatively recent, it was important to search not only electronic bibliographic databases/ journals but websites of research centers as well. Efforts were made to include the grey literature – i.e., unpublished studies, such as those found through the World Bank and the Abdul Latif Jameel Poverty Action Lab at MIT, as well as other organizational publications. Citation tracking, searching conference presentations, and contacting researchers in this field (as well as examining the body of work of the main influential authors in this field) also proved to be integral to the identification of impact evaluations meeting the criteria described above. The search was conducted from June 1, 2012 – April 9, 2013. Papers in press were updated with their most recent versions post-April, 2013 when applicable.

Electronic database search: Searches were conducted using 1.) "meta" search engines (containing results from multiple databases), 2.) individual databases in education and economics, and 3.) individual journals in these fields (which are known to publish education impact evaluations in Africa). Regarding the "meta" search engines, I conducted a Columbia University "custom search" which pulled studies from all databases in the fields of Social Sciences, Economics, International/Area Studies, and Political Science/ International Affairs. Secondly, I searched the following eight databases individually in order to improve the precision of my results (these databases allow searches using additional database-specific "subject terms" which cannot be used if all databases are being searched at once): JSTOR, Academic Search Complete, Business Source Complete, EconLit with Full Text, Education Full Text (H.W. Wilson), Education Research Complete, ERIC, and Social Sciences Full Text (H.W. Wilson). Finally, I conducted a well-documented electronic search of 25 individual journals. Please see Appendix B for a full list of all databases, journals, and websites searched. Upon request, I can provide documentation of the search terms, date of search, dates under search, and results for each search string in every database, journal, and website.

Search Terms: Regarding the search term strategy, the search terms were adapted for each database/ meta-search and varied in complexity depending on the limitations of each search engine. In all cases, the "education" and "region"- specific search terms were used. When possible, all or portions of the "methodological" and "intervention area" search terms were used as well. Again, when possible, each country was searched for individually. A full list of search terms can be found in Appendix B.

Additional search strategies: In addition to the electronic databases searches, I conducted a well-documented search of 22 organizational websites. As these websites permitted "education only" filters, I examined all results found under this filter. Further, I employed "citation tracking" methods for studies/ books/ lit reviews that were found to be particularly influential and comprehensive and were thus searched systematically for impact evaluations in Africa (Appendix B). In addition, I contacted researchers themselves and searched authors' websites and curricula vitae for additional citations.

Finally, I searched relevant conference websites for presentations and papers. I tracked references from three major World Bank conferences pertaining to my research questions. Please see Appendix B for the complete list of organizational websites searched, studies used for citation tracking, individuals contacted, and researcher as well as conference websites searched. Any working paper/ mimeo/ draft reference found through these methods was then entered into Google Scholar to check that it had not been published recently.

Results of literature search: The results of my systematic search are found below (Table 4). Out of 10,660 citations, I identified 168 articles that appeared to fit the criteria outlined in Section B (above), based on their abstracts alone. After a more thorough read of these 168 articles, 112 additional had to be further eliminated, namely due to the fact that a.) their findings could not be standardized with the information given in the paper, b.) the methodologies employed were not truly experimental or quasi-experimental, or c.) they were missing learning-specific outcomes (for example, often abstracts would state that the intervention had an impact on educational outcomes but would not state at the outset whether these were enrollment, attendance or learning outcomes, etc.). In total, this meta-analysis includes 56 articles which contain 66 separate studies or experiments (e.g., some articles contain experiments conducted in two separate countries), corresponding to 83 treatments (some experiments contain multiple treatment arms), and 420 effect size estimates.

Table 4. Results of systematic search process

Total relevant articles after following the search process above:	**10,660** (+ citation tracking etc.)
Total articles meeting criteria (based on abstract alone): *After a more thorough read of these 168 studies:*	168
➢ No. without an experimental or quasi-experimental design	-33
➢ No. without achievement outcomes	-31
➢ No. outside of SSA	-2
➢ No. w/o explict findings	-6
➢ No. duplicates (earlier dates)	-18
➢ No. w/ non-school-aged population	-5
➢ No. whose findings can't be standardized	-17
Total number of articles to be included in learning outcomes meta-analysis:	**56 articles** (containing 66 experiments & 83 treatments)

D. Coding of studies and effect size choice/calculations

Using the studies collected under the systematic literature search, I then applied a 74 question coding sheet that I developed to each study in order to extract information on key variables (see Appendix I for coding sheet). I then subsequently entered these values into a Microsoft Access database.

Background and moderator variables: I coded background and moderator variables that I considered would be useful in the analysis portion of this meta-analysis (to test for heterogeneity or for sub-group analysis for example). Examples of these variables include the following: country, region within Africa, setting (rural/ urban), length of treatment, publication type, academic field, methodology, assessment type, whether or not the assessment was psychometrically evaluated, and subject (math, language). I also collected data on the relative influence of the peer reviewed journals in which these studies are found; I recorded both the Scientific Journal Ranking metric[6] (accounts for journal prestige as well as frequency of citation), as well as the Article Influence score (AI), which measures the average influence of articles found in a particular journal (this measure is related to the Eigenfactor score).[7]

Creation of quality index: While the main inclusion criteria listed in Section B (above) restrict studies to only those with the most rigorous methodologies (experimental and quasi-experimental), quality variation in studies still remain. I thus created a quality index to use as a way to control for study quality in sensitivity analyses. This index pulls from both the Cochrane risk of bias framework (Higgins et al, 2011) and the GRADE system for method quality (Higgins and Green, 2011) when their use is applicable to this literature. This index evaluates both the clarity of study/ intervention design and the integrity of the identification strategy used. For most methodologies used in this set of studies, the index includes basic metrics such as clarity of the study design/ intervention and balance or overlap at the outset on observable characteristics of treatment and control groups. However, additional metrics are added for differing study methodologies. For example, for randomized controlled trials I also include sample attrition as a metric, while for instrumental variables studies, I additionally evaluate whether or not the first stage regressions were evaluated, as well as if the authors conducted tests of the exclusion

[6] http://www.scimagojr.com/journalrank.php

[7] http://www.eigenfactor.org/

restriction. The quality index is on a scale of 1-6, regardless of methodology type (see Appendix D for further details).

Effect size choice and calculations: The goal of meta-analysis is to pool and compare the outcomes of a population of studies on a particular topic. Since studies collect different outcomes on different scales, in order to make comparisons, these must be standardized. In meta-analysis, there are three common families of effect sizes that are used for standardization: the d-family, the log odds family, and the correlation family. For this meta-analysis, I used Cohen's *d* as my standardized effect size (see below). The standardized mean difference is commonly used in experiments and is used in situations in which the independent variable is binary and the dependent variable is continuous. In general, if we assume that in the treatment group the outcomes Y_{it} ~N(μ_t, σ^2), and in the control group, Y_{ic} ~ N(μ_c, σ^2), then the standardized mean difference is defined as,

$$\delta = \frac{\mu_t - \mu_c}{\sigma}$$

where μ_t is the average outcome in the treatment group, μ_c is the average outcome in the control group, and σ^2 is the common variance. In single level experiments (with no clustering), δ can be estimated using the sample means and variances, as Cohen's d, where

$$d_{sm} = \frac{\bar{y}_T - \bar{y}_C}{s_{pooled}},$$

and

$$s_{pooled} = \sqrt{\frac{(n_T - 1)s_T^2 + \ (n_C - 1)s_C^2}{n_T + n_C - 2}}$$

are a function of the sample means ($\bar{y}_T$ & $\bar{y}_C$), the pooled sample standard deviation standard deviation (s_t), and the sample sizes (n) for each group (treatment and control) (Lispsey and Wilson, 2001). Importantly, when $n_t = n_c = n > 20$, Cohen's *d* is an unbiased estimator of δ.

In Africa, many large-scale impact evaluations take place in multiple sites. In these situations, the units randomized to the treatment conditions are clusters, for example schools, villages, or classrooms. In this case, as Hedges (2007) points out, it is no longer simple to define σ^2, since the total variation in the outcomes involves two components: the variation within-clusters (σ_w^2) and the variation between clusters (σ_b^2). It is common in econometrics to standardize the outcome variable before conducting analyses, thus making the δ_t effect size the most appropriate, where:

$$\delta_t = \frac{\mu_t - \mu_c}{\sigma_t}$$

where $\sigma_t^2 = \sigma_w^2 + \sigma_b^2$. Importantly, this is not the same effect size as is commonly reported in single-site small scale experiments (as above). In Mathematical Appendix C., I provide the Hedges (2007) method for relating these two, and throughout this meta-analysis I take care to convert all effect sizes to one estimating δ_t instead.

An important problem encountered in this meta-analysis is that in different academic disciplines, study findings are presented very differently. For example, the majority of studies within the field of economics present their findings in regression tables (with accompanying robust standard errors), while many authors of studies published in education journals report their findings in terms of raw means (and standard deviations) or gain scores (and standard deviations), often alongside a one-way ANOVA (with two or more groups), a two-way ANOVA, or an ANCOVA table. I thus follow statistical procedures outlined in Lipsey and Wilson (2001), Appendix C, Hedges (2007), and Tipton (2014) to extract comparable effect sizes from these disparate measures.

In a number of studies from the education literature, the reported standard errors and variances do not take into account the actual clustering in the data. For example, multiple schools are assigned to treatment conditions, but the analysis treats the students as independent. In order to correct the outcomes for this clustering, I follow results found in Hedges (2007) (see Mathematical Appendix C.). These results require information on the degree of clustering (the intra-class correlation, ICC) found in the data, which is generally not reported. Instead, following common recommendations, I impute this information using ICC calculations from TIMMS and SACMEQ estimates. When available, I use country and subject-specific data (Postlethwaite, 2004 and Zopluoglu, 2012).

In a handful of cases, economics papers provided an unstandardized effect size (in a regression table) along with robust or cluster robust standard errors, but no standard deviation information (Martinez, Naudeau, and Pereira, 2012; Lassibille, Tan, Jesse, and Nguyen, 2010; Kazianaga, de Walque, and Alderman, 2012). In these cases, I was able to extract a standardized effect size using derivations by Tipton (2014), based upon Hedges (2007) (Mathematical Appendix C.).

Finally, in one study the outcome reported was based on an instrumental variable analysis (Reinikka and Svensson, 2011). The findings were reported in a regression table in which the independent variable was continuous (an increase in 1 percentage point in the share of funding going to schools (due to information campaign) is associated with an certain increase in standardized test scores). This type of effect size is from the *r*-family (correlational effects) of effect sizes. In order to put this estimate on the same scale as the experiments in the rest of my sample, I dichotomize this variable following Lipsey & Wilson (2001) and Tipton (2014) (see Appendix C). The resulting effect size can be interpreted as the impact of the full share of funding reaching primary schools (due to the information intervention) versus a scenario in which none of the funding reached the schools.

Unfortunately, 17 studies whose findings could not be converted to standardized effect sizes were excluded. An example of a study that could not be included is a study on the impact of pupil-teacher ratios in South Africa (Case and Deaton, 1999). In this case, information with which to standardize the findings and dichotomize the continuous treatment variable (resulting from an instrumental variables methodology) is not available. Other examples of excluded studies include certain papers from the education literature which provide findings in the form of a two-way ANOVA or an ANCOVA table without any additional information on the group means or on the correlation structure between the covariates and dependent variables (for ANCOVAs) (examples include: Awolala, 2011; Onabanjo and Okpala, 2006; Adedayo, 1999; Kurumeh, 2008; Olowa, 2009, and Okoye, 2010). Table 4 above provides a full account of the elimination process.

Conventions for recording of effect sizes: It is also important to note that many studies report multiple effect sizes due to measures on multiple grades, in multiple subjects, for multiple sub-groups, or for multiple time periods. I record all relevant effect size estimates and control for any inter-study dependencies (students across grades or subjects) with robust variance estimations (see Section E below).

In addition, while I record effect size estimates for all subject, grades, and sub-groups, I use the effect size associated most closely with the intervention when I conduct the overall meta-analysis. For example, if a program targeted reading methods and also happened to test students on their mathematics performance post-intervention, I use the results most closely tied to the intervention (language scores) as the main outcome of interest in the main analysis. Also note that I recorded the regression table and column/ row from which each estimate came.

Inter-rater reliability of effect sizes reported: In meta-analysis, one method for reducing bias is through repeated coding of the same set of studies to ensure that there is a high degree of reliability between these measurements. I measured reliability in two ways. First, I re-entered a random subset of approximately 40% of studies (both effect size measures as well as moderator variables) approximately three months after the first data collection. I found that 96% of the effect sizes recorded in round 1 matched those entered in round 2. Secondly, an independent researcher re-entered a random subset of effect sizes (approximately 51% of estimates), and the inter-rater reliability estimate was 98%. The few discrepancies were discussed, and the final effect size estimates were agreed upon mutually.

E. Selection of Estimation Model

In meta-analysis, there are two models commonly used for pooling effects. The first is a "fixed effects" analysis, in which it is assumed that every study is estimating one common or "true" effect. In contrast, in "random effects" analyses, it is instead assumed that each study aims to estimate a different effect and that the meta-analysis includes a distribution of "true" effects. Since I am pooling results from a variety of interventions and regions within Africa, throughout I use a random effects model, since it assumes and accounts for heterogeneity. I further employ robust variance estimation methods with small sample corrections to control for inter-study dependencies (Hedges, Tipton, and Johnson, 2010; Tipton, *in press*).

Traditional random effects model: In this dissertation, I use random effects models with subgroup analyses and meta-regression to account for heterogeneity in effect sizes. To estimate the combined effect size under a random effects analysis, the individual study-specific effect sizes are weighted by their inverse variance (and in addition random effects are added to account for heterogeneity between studies). The equation for the weighted average effect size using a random effects model is as follows:

$$\bar{T} = \frac{\sum_{i=1}^{m} w_i T_i}{\sum_{i=1}^{m} w_i}$$

where the optimal variances are $w_{\mathrm{i}} = 1/\mathrm{V}(T_{\mathrm{i}}) = 1/(v_{\mathrm{i}} + \tau^2)$ where V(.) is the variance. In this case, the sampling error of the average effect size can be calculated using,

$$V(\bar{T}) = \frac{1}{\sum_{i=1}^{m} w_i} = \frac{1}{\sum_{i=1}^{m}\left(\frac{1}{v_i + \tau^2}\right)}$$

where $\bar{T}$ stands for the combined effect size, w_{i} is the weight given to study *i*, *m* is the number of studies, and v_i is the within-study variance of study *i* (a known value), and τ^2 is a measure of between study heterogeneity estimated using a method-of-moments estimator. Note that this model assumes that the effect sizes are independent. If indeed the effect sizes are not independent, this dependency should either be modeled or addressed using robust variance estimation, as described below.

In random effects, it is assumed that the true effect sizes vary from study to study. In addition to estimating an average effect, it is therefore important to also estimate the degree of heterogeneity. Here the distribution of true effect sizes is estimated as having a mean of $\overline{T}$ and a variance of τ^2. Since variances are not easy to compare, a standardized measure of variance, I^2, is commonly used. I^2 is a function of τ^2 and represents the proportion of the total variation between effect size estimates attributed to real study differences.

Meta-regression: In situations in which the effect sizes are heterogeneous, one method for exploring this heterogeneity is through meta-regression (Lipsey and Wilson, 2001). Meta-regressions are modified weighted least squares regressions with the effect size as the dependent variable, while study descriptors or moderators (e.g., study population, methodology, region) can be entered as independent variables in order to ascertain whether or not a portion of the heterogeneity can be explained by known study features. The RE meta-regression model can be written as follows:

$$T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p \beta_{ip} + \phi_i + \epsilon_i$$

$$\epsilon_i \sim \mathrm{N}(0, v_i)$$
$$\phi_i \sim \mathrm{N}(0, \tau_c^2)$$

In this model, I assume that the effect sizes T_i can be explained as a function of a set of p covariates, and that there may be some residual unexplained true variation (τ_c^2) and sampling error (v_i). Here $\tau_c^2 < \tau^2$, since some proportion of the heterogeneity has been explained by these moderators. Note that it is standard to assume that the ϕ_i are independent of the ϵ_i. These regression coefficients can be estimated using weighted least squares, where the weights are typically chosen to be inverse variance (e.g., $w_i = 1/(v_i + \tau_c^2)$) for efficiency purposes. The variance component (τ_c^2) is estimated using a method-of-moments estimator, and the within-study variances (v_i) are assumed to be known.

Robust Variance Estimation: In this meta-analysis, many studies report multiple outcomes. For example, the same units (students) might be measured multiple times in different subjects or, one study may involve multiple treatment arms and only one control group (thus students in the control group are measured multiple times). In some cases, a single study may report the results of multiple independent experiments in different countries or contexts (thus methodologies or estimation strategies may overlap). Traditional meta-analysis models, however, have required that the effect sizes included in a meta-analysis model are independent, which is clearly not the case.

A recent solution to this dependence problem is to use robust variance estimation (Hedges, Tipton, and Johnson, 2010). The robust variance estimation approach (RVE) is similar to that of cluster robust standard errors used commonly in econometrics and is a type of Taylor series or linearization estimator (Tipton, *in press*). Importantly, the RVE approach does not require the same distributional assumptions and does not require the use of inverse-variance weighting. However, since inverse-variance weighting leads to more efficient estimates of both the average effect size and moderators, in this dissertation I use approximately inverse variance weights, as defined in Hedges et al.

Importantly, when the number of studies in a meta-analysis is small (fewer than 40 independent experiments), it has been shown that RVE procedures can result in invalid standard error and p-value estimates. For my full sample analysis, I have a large sample (a total of 66 experiments), but for topic-specific analyses, the number of studies within each intervention area is much smaller. Regardless, I use small-sample corrections throughout my analysis (Tipton, *in press*), which adjust both the residuals used in the RVE and the degrees of freedom of the estimates. These degrees of freedom depend on both the number of studies included in the meta-analysis and features of the particular covariate under study. When these degrees of freedom are smaller than four, simulation results indicate that the p-value associated with estimates should not be trusted. These small-sample correlations are particularly important when conducting topic-level "mini-meta-analyses."

Throughout this dissertation, I therefore account for dependence in the following ways. First, I treat independent experiments reported within the same article as separate studies. This means that there are at total of 66 studies included in this analysis. Second, each of these studies reports between 1 and 12 effect sizes/ outcomes. In order to estimate the average effect size overall, for subgroups, and in meta-regression models, I use RVE with the "correlated effects" weights. I use an assumed correlation of $\rho = 0.80$, and where necessary, conduct sensitivity analyses. (This is the procedure recommended by Hedges et al). I estimate these models in the statistical program **R** using the **robumeta** package (Fisher, 2014).

CHAPTER V.

RESULTS

This chapter is organized around the three main research questions in this meta-analysis (described fully in Chapter III). Part A describes the state of the literature, including research gaps and other descriptive statistics of studies in this sample. Part B provides an overview of relative pooled effect size measures and tests the robustness of these relative findings. Part C explores variation in effect size estimates across the full sample, as well as by intervention type. Finally, in Part D, I test for publication bias among the studies in this sample.

A. What is the state of the literature in this area?

I first present my results in regard to the state of current research in educational effectiveness in Sub-Saharan Africa. I examine research gaps by location and topic, as well as highlight a series of descriptive statistics about this selection of literature.

1- Research gaps by location and topic

One important finding of this meta-analysis is that the number of studies taking place in Sub-Saharan Africa is much larger than that of previous reviews, which contained between 9 – 19 studies each (Kremer and Holla, 2009; Kremer, Brannen, and Glennerster, 2013, and McEwan, 2013). The studies are relatively evenly spread out by region (across West, East, and Southern Africa), but there are certain countries in which a disproportionate bulk of the research has been conducted (Kenya, Nigeria, South Africa, Burkina Faso, Uganda, and Madagascar) (see map of research availability in Figure 1, Appendix E). In particular, over one-third of the studies in this sample were conducted in Kenya. Of these, 68% are from the field of economics, 10% from education, and 22% from public health. Further, there are a number of countries in which no experiments have been conducted at all (Zimbabwe, Angola, Gabon, and Mauritania, among others). A number of these countries are affected by war or political instability, but this is not the case for all countries with no previous research.

Regarding topical coverage, please see Table 5 below for a summary of the number of studies by intervention area. Note that only studies of interventions are included in my main analysis. Thus while I collected studies which examine the relative advantage (or disadvantage) of attending the private school system, I do not include these studies in my analysis of the average intervention-level effects. This is due to the fact that such studies do not entail a program or policy shift but instead capture differences due in part to the result of the education marketplace general equilibrium. However, I do briefly analyze this set of studies separately at the end of Chapter V (Section C, Part 2).

When I examine the results of my search by intervention area (Table 5 below), I find that there is a dearth of evidence in many of the fields under study, yet a reasonable amount of research in others (e.g., school health-related interventions and pedagogical interventions). This fact limits what type of analysis I can conduct within (or across) certain intervention areas but does shed light on my research question regarding what topics are understudied in this is literature. Further, while the categories (below) represent those for which I searched, some of them are aggregated later for analysis purposes. For example, “Language of Instruction” and “Technology-Assisted Learning” are folded into the “Instructional Interventions” category, while “Abolishment of School Fees” and “Cash Transfers” are both folded into a “Cost Reduction” category.

Table 5. Results of systematic search (by topic)

INTERVENTION AREA	No. of Experiments
After-school Tutoring (non-computer-based)	0
High Stakes Testing & Accountability Systems	0
School Choice	0
Abolishment of School Fees	1 - 2
Class Size	1 - 2
Instructional Time	1 - 2
Language of Instruction	1 - 2
Student Incentives	1 - 2
Tracking & Peer Effects	1 - 2
Cash Transfers & Free Uniforms	3 - 6
Information for Accountability	3 - 6
Infrastructure + Complementary Inputs	3 - 6
Management Interventions	3 - 6
School Meals & Supplements	3 - 6
School Supplies Provision	3 - 6
School Type (Private Schools)*	3 - 6
Technology-Assisted Learning	3 - 6
School Health	7 - 12
Teacher Incentives	7 - 12
Pedagogical Methods	12+

**Studies of "private school" advantage were removed from the main analysis as do not entail a program or policy shift but instead capture differences due in part to the results of the education marketplace general equilibrium.*

The lack of evidence in certain areas such as instructional time or class size is surprising given, for example, the number of schools in Sub-Saharan Africa that use multi-grade or multi-shift classrooms or that have classrooms of over 100 students (Brown, 2010). Further in 1992, Kellaghan and Greaney examined ways in which national examinations and accountability systems can be used to improve student learning throughout fourteen African countries, yet the hypotheses they outline in their paper have also yet to be evaluated (there are no studies of high-stakes testing etc.). In addition, Selod and Zenou (2003) run a number of simulations models trying to predict the optimal type of voucher program in South Africa, but their findings have also yet to be tested empirically. Finally, there are no supplemental, after-school tutoring programs evaluated in this literature, despite the fact that such programs have been very successful in other developing countries (Banerjee et al, 2007).

2- Descriptive statistics of full sample

Table 6 below describes the data in this sample overall, as well as by academic field. Overall, 62% of the studies in this sample are from the field of economics, while 23% are from the field of education, and 15% are from public health. Descriptive statistics of note include the fact that studies in the public health field are much more likely to use tests of cognition than any other field (53.77% use cognition tests), and that studies in the economics literature are much more likely to employ only "composite score" measures (combining academic subjects) than the fields of education or health. Regarding region, studies in the fields of public health and economics are more common in East Africa, while studies within the education literature have a larger focus in West Africa. In addition, studies from the economics field are more likely to be nationally representative (35%), versus 7% from the field of education and 0% from public health. Other differences of note include the fact that papers in the fields of education and health are predominately from peer-reviewed journals (~70%), while those in the field of economics are more likely to be working papers (over 50%). The vast majority of studies in all fields were published after 2000 (100% for economics, 100% for education, and 90% for public health).

In terms of study methodology, education and health studies employ predominantly experimental methods, while a reasonable number of papers in the field of economics employ quasi-experimental methods (30% of studies). Regarding assessments, studies in the field of economics are more likely to be researcher designed (56% for economics versus 47% in education and 40% in health). Moreover, a very small fraction of those examinations in the economics field are subject to psychometric analysis (2.44%, compared to 80% for education studies and 50% for public health studies); this may be detrimental for studies in the field of economics if those examinations lack discriminatory power. Regarding the size of the experiments, studies in the field of economics and health tend to be larger than studies in the field of education (in terms of randomized units), and studies in the economics literature are the largest in terms of the number of schools included. In addition, all fields are equally unlikely to report scores for follow-up/ retention after the end of the intervention (approximately 10% of studies from each field report these follow-up outcomes). And finally, the average length of interventions differs by field, as studies in the education literature last on average 7-8 months, while those in the health and economics literature last a year or more.

Table 6. Descriptive statistics (full dataset)

	All	Economics	Education	Health
MODERATOR VARIABLES	N=66	N=41 (62.12%)	N=15 (22.73%)	N=10 (15.15%)
SUBJECT (%)				
Cognition	11.64	2.06	0.00	53.77
Composite	22.84	45.88	2.25	12.26
Language	36.64	26.80	59.55	19.81
Math	23.28	24.23	25.84	11.32
Science	3.02	0.00	10.11	0.00
Social Science	2.59	1.03	2.25	2.83
REGION (%)				
East Africa	48.48	58.54	13.33	69.00
Southern Africa	22.73	17.07	33.33	30.00
West Africa	28.79	24.39	53.33	10.00
SETTING (%)				
Rural	42.42	46.34	6.67	80.00
Urban	4.55	2.44	13.33	0.00
Both rural & urban	28.79	41.46	13.33	0.00
Setting not reported	24.24	9.76	66.67	20.00

[Table 6 continued on next page].

Table 6 (continued)

	All	Economics	Education	Health
MODERATOR VARIABLES	N=66	N=41 (62.12%)	N=15 (22.73%)	N=10 (15.15%)
PUBLICATION TYPE (%)				
Peer-reviewed journal	53.03	41.46	73.33	70.00
Report	13.64	7.32	26.67	20.00
Working paper	33.33	51.22	0.00	10.00
IDENTIFICATION STRATEGY (%)				
Difference-in-Difference	3.03	4.88	0.00	0.00
Changes-in-Changes	1.52	2.44	0.00	0.00
Instrumental Variables	1.52	2.44	0.00	0.00
Matching (simple)	3.03	0.00	13.33	0.00
Matching (non-parametric)	6.06	9.76	0.00	0.00
Matching (propensity score)	3.03	4.88	0.00	0.00
Natural Experiments	1.52	2.44	0.00	0.00
Randomized Controlled Trial	77.27	68.29	86.67	100.00
Regression Discontinuity	1.52	2.44	0.00	0.00
Time Series	1.52	2.44	0.00	0.00
LEVEL (%)				
Primary through Secondary	1.52	2.44	0.00	0.00
Primary	83.33	92.68	53.33	90.00
Secondary	15.15	4.88	46.67	10.00
ASSESSMENT TYPE (%)				
Adapted standardized	22.73	12.20	26.67	60.00
Researcher designed	51.51	56.10	46.67	40.00
Standardized	25.76	31.71	26.67	0.00
PSYCHOMETRICS REPORTED (%)	27.27	2.44	80.00	50.00
NATIONALLY REPRESENT. (%)	22.73	34.15	6.67	0.00
RANDOMIZED UNITS>30 (%)	86.36	95.12	60.00	90.00
NUMBER OF SCHOOLS>=30 (%)	74.24	95.12	40.00	40.00
TEST OF BALANCE (if applies) (%)	86.67	91.43	80.00	80.00
RESULTS MIDLINE (%)	0.08	0.07	0.00	0.20
RESULT POST-ENDLINE (%)	0.14	0.15	0.13	0.10
PUBLICATION POST-2000 (%)	98.48	100	100	90
PUBLICATION POST-2010 (%)	71.21	80.49	66.67	40
PROGRAM LENGTH (mean in months, SD)*	15.0 (9.977)	19.15 (8.703)	7.75 (10.238)	11.83 (5.292)

*These averages were calculated without data that represented follow-up tests occurring over three years later.

B. What is the relative effectiveness of different intervention types?

This chapter examines the overall effect of various education interventions in Sub-Saharan Africa and explores differences among the 12 intervention types included in this paper. The analyses include the following sample restrictions: For my main analysis, I remove studies that examine private school advantages due to the fact that these studies only examine the general equilibrium result of a private versus public school system and do not entail an intervention or actionable program/ policy. I will however, briefly discuss these studies at the end of section C. Further, I restrict the sample to only academic or "curricular" outcomes; effects on cognitive outcomes (memory and attention) are restricted to school health-related studies (and one infrastructure study) and are analyzed separately within those intervention areas. Finally, unless otherwise noted, the effect sizes used in the main meta-regressions here are for all students (not sub-groups of students) and are measured at the end of the intervention period.

1- Overall effect size and heterogeneity statistics

While I am more interested in comparing pooled effects sizes between topics/ intervention areas, I first estimate a pooled effect size across all intervention areas for comparative purposes. In this sense, I am estimating the mean impact of education interventions in Sub-Saharan Africa on students' academic outcomes (not cognitive outcomes). When I estimate the random effects model (see Table 7, below), the overall effect size (or pooled estimate) is 0.181 standard deviations, with a 95% confidence interval of (0.091, 0.27). The pooled effect size is very precisely estimated (SE = 0.045, df = 47.8, p = 0.00018). Regarding measures of heterogeneity, τ^2 is estimated to be 0.036, and the variation in effect size attributable to heterogeneity (and not sampling variation), I^2, is 89.32%. This I^2 value is quite high (and relatively rare), but I expect such a degree of heterogeneity with this data given that I am pooling evaluations of studies using different types of interventions.

Table 7. Overall pooled effect size estimate

	Estimate	SE	t-value	df	(P|t|)	95% CI.L	95% CI.U
Pooled ES	0.181***	0.045	4.06	47.8	0.00018	0.091	0.27
Number of studies = 60							
Number of outcomes = 138 (min = 1, mean = 2.3, median = 2)							
ρ = 0.8; I^2 = 89.32; τ^2 estimate = 0.036							

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

2- Pooled effect sizes & heterogeneity for each intervention area

Next, I examine the relative impact of these different intervention areas. The pooled effect sizes range in magnitude from a statistically insignificant -0.008 standard deviations for health treatment interventions to an extremely high estimate of 0.918 standard deviations for pedagogical interventions. Table 8 below displays the results of this meta-analysis by intervention area (with heterogeneity statistics). However, this list is not meant to be at all definitive in its ranking of intervention areas. While there are some intervention areas whose results are both statistically significant and have high degrees of freedom (pedagogical interventions), most of the robust variance weighted averages in this list have too few degrees of freedom to result in significant pooled effect sizes. Since the RVE method does not perform well in terms of hypothesis testing with fewer than four degrees of freedom, I include the estimated effect sizes but not the associated p-values. In these cases, the results should be interpreted descriptively.

Within an intervention area with multiple entries, I would expect to find the heterogeneity statistics to be lower (on average) than those for the entire pooled sample, as the intervention types are becoming much more comparable. For example, within the Management Intervention category (this includes studies in school-based management or district/ sub-district-level management), the overall effect size is equal to an statistically insignificant 0.016 standard deviations (SE = 0.028, df = 3.14), however the variation in ES attributable to heterogeneity between studies (I^2) is reported to be only 18.7%, which is smaller than most other I^2 estimates, meaning there is less true variation between studies. However, even within certain intervention areas, studies may still vary by intervention characteristics, populations, or other contextual factors, which contributes to the high degree of heterogeneity (I^2) observed within a number of these intervention areas (notably: infrastructure with complementary inputs, teacher incentives, and pedagogical interventions). In some cases, due to small sample size, the I^2 estimate is zero.

Table 8. Pooled effect size measures by intervention area

		Std.			95%	95%		
	Estimate	Error	df	(P|t|)	CI.L	CI.U	I^2	t^2
Quality of Instruction								
School Supplies Provision	0.022	0.015	1.25	‡	-0.10	0.14	0	0
Class Size & Composition	0.109	0.049	1	‡	-0.51	0.73	0	0
Instructional Time†	0.464	0.198	…	…	…	…	…	…
Pedagogical Intervention	0.918**	0.314	15.1	0.01	0.25	1.59	95.6	0.
Student Cognitive Ability								
Health Treatment	-0.008	0.041	3.5	‡	-0.13	0.11	45.7	0
School Meals/Supplements	0.059	0.022	1.16	‡	-0.15	0.27	0	0
Student/ Teacher Motivation								
Student Incentives	0.288	0.015	1	‡	0.10	0.48	0	0
Teacher Incentives	0.075	0.047	4.89	0.2	-0.05	0.20	53.45	0
School/ System Accountability								
Management Intervention	0.016	0.028	3.14	‡	-0.71	0.10	18.72	0
Information Provision	0.147	0.053	1.63	‡	-0.14	0.43	0	0
Financial Limitations								
Cost Reduction Intervention	0.036	0.036	1.58	‡	-0.16	0.24	27.3	0
Infrastructure + Add. Inputs	0.189	0.122	1.97	‡	-0.35	0.72	94.23	0.
OVERALL	0.181	0.045	47.8	0	0.09	0.27	89.32	0

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

† There was only one study which focused on the impact of an increase in instructional time.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

3- Robustness tests of high and low impact intervention types

In this section, I examine the robustness of the pooled effect size results above. Specifically, I test whether the apparent effectiveness of interventions in pedagogical methods or the relative low performance of interventions in school health programs hold within restricted samples, as well as in meta-regression models.

Effectiveness of pedagogical methods: I first examine the large average effect size (0.918 standard deviations) associated with pedagogical interventions; this effect size is significant at the 1% level, though there is a significant amount of heterogeneity ($I^2 = 95.65$) within this intervention category. This category includes interventions in which the method of instruction or teacher pedagogy was altered, as well as interventions that employ the use of blended/ technology-assisted learning. To explore this further, I first test whether pedagogical interventions have a higher effect size on average than other interventions in the full sample; I find that instructional interventions have on average an effect size 0.540 standard deviations higher than any other intervention type, significant at the 5% percent level (df = 13.9, p = 0.025).

Table 9. Relative impact of pedagogical interventions (unadjusted model)

	Estimate	Std Err	t-value	df	(P|t|)	95% CI.L	95% CI.U
Intercept	0.0807***	0.0197	4.1	37.3	0.00022	0.0408	0.121
Pedagogical	0.5418**	0.2149	2.52	13.9	0.02452	0.0807	1.003

Number of studies = 60
Number of outcomes = 138 (min = 1 , mean = 2.3 , median = 2 , max = 9)
$\rho = 0.8$
$I^2 = 88.19$
τ^2 estimate = 0.033

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

This differential is of quite a large magnitude, given that most effect sizes of educational interventions often have effect sizes closer to 0.20 standard deviations (Hill, 2008). However, the quality of studies in this intervention area are on average lower than those in non-pedagogical categories due largely to the fact that studies in this area are slightly less likely to report multiple balance statistics on observable characteristics of students (across treatment and control groups at baseline) and are also less likely to report or address student attrition. Scores for a subset of these studies thus receive lower rankings on the quality index, described in Chapter IV. In addition, there are two studies with effect sizes over three standard deviations (see the forest plot in Figure 2, Appendix F.), and while these effect sizes have large confidence intervals, examining the pooled effect size without these outliers will be important. Finally, as studies in this intervention area employ both matching and randomized controlled trial identification strategies, I ask whether methodological differences could account for these large effects.

To explore these concerns further, I first restrict the sample of pedagogical studies to the following sub-samples: randomized controlled trials only, the sample without outliers, randomized trials only (without outliers), and high quality studies only (quality index greater than or equal to three). In Table 10 (below), I find that various sample restrictions do greatly affect the magnitude of the pooled effect size but that this effect size remains relatively high and statistically significant, even when restricting the sample to only the highest quality studies (*d* = 0.228, SE = 0.078, df = 5.2, p = 0.032).

Table 10. Sample restrictions/ sensitivity of pooled effect size (for pedagogical methods)

Restricted Samples	Estimate	Std. Err.	df	P(|t|>)	95% CI.L	95% CI.U	Studies
Full sample	0.918**	0.314	15.1	0.0104	0.249	1.59	17
RCTs only	1.000**	0.356	13.1	0.0145	0.233	1.77	15
No outliers	0.536***	0.172	12.9	0.0825	0.164	0.908	15
RCTs, no outliers	0.566**	0.194	11	0.0141	0.138	0.993	13
High quality	0.228**	0.078	5.2	0.0317	0.029	0.426	9

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

In addition to sample restrictions, I also employ meta-regression to test the robustness of the differential impact of changes in pedagogical methods (see Table 11 below). Addressing some of the concerns above, I employ moderator variables controlling for the following: the quality of the study, the identification strategy employed, subject matter of the assessment, the level of the student population, the publication type, region, whether the study is nationally representative, the length of time of the intervention, and the assessment itself (whether the psychometric properties of the assessment were measured or whether the assessment was researcher-created or a standardized exam (versus a reference category of an adapted standardized test)). I find that controlling for these factors, the average differential impact of pedagogical methods is slightly above 0.30 standard deviations (see Table 11 below). And when I conduct this analysis on the sample without outliers, I find that the coefficient decreases to 0.257, but it is still significant at the 5% level (SE = 0.112, df = 9.3, p = 0.047) (see Table 26, Appendix G).

Overall, I find that the pooled effect size associated with these pedagogical interventions to be relatively robust to both the addition of moderator variables and to various sample restrictions. And when the sample is limited to only high quality studies, the pooled effect size associated with pedagogical interventions drops considerably but is still highly significant and large in comparison to other intervention types in this sample.

Regarding why these interventions are estimated to be so effective, it is possible that the quality of current pedagogical methods is on average so poor (and current learning levels so low) in much of Sub-Saharan Africa, that even small changes in teaching techniques could have a large impact. This would be in keeping with the Heyneman-Loxely hypothesis (1983), based on these authors' cross-country analysis, which posits that in low-income countries, "the predominant influence on student learning is the quality of the schools and the teachers to

Table 11. Average differential impact of pedagogical interventions (adjusted model)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	0.081***	1.235*	1.548*	1.546*	1.740*	1.738	2.035
	(0.000)	(0.095)	(0.059)	(0.087)	(0.093)	(0.106)	(0.134)
Pedagogical	0.542**	0.302**	0.330**	0.326**	0.331**	0.324**	0.321*
	(0.025)	(0.041)	(0.027)	(0.027)	(0.033)	(0.009)	(0.054)
Assessment		-0.032	-0.020	-0.036	-0.009	-0.011	0.028
(researcher)		(0.641)	(0.825)	(0.722)	(0.931)	(0.914)	(0.801)
Assessment		0.002	0.075	0.086	0.205	0.266	0.399
(standardized)		(0.992)	(0.681)	(0.651)	(0.412)	(0.322)	(0.248)
Language		-0.059	-0.076	-0.068	-0.077	-0.114	-0.052
		(0.389)	(0.348)	(0.492)	(0.468)	(0.225)	(0.555)
Math		-0.143	-0.136	-0.128	-0.126	-0.155	-0.180
		(0.137)	(0.229)	(0.294)	(0.339)	(0.222)	(0.189)
Science		0.556	0.350	0.358	0.234	0.235	-0.044
		(0.519)	(0.694)	(0.696)	(0.804)	(0.804)	(0.966)
Soc. Science		0.534	0.571	0.626	0.663	0.635	0.562
		(0.731)	(0.716)	(0.702)	(0.683)	(0.702)	(0.728)
Reliability		-0.246	-0.187	-0.153	-0.209	-0.179	-0.159
(of assessment)		(0.286)	(0.397)	(0.613)	(0.546)	(0.669)	(0.720)
Quality		-0.192	-0.180	-0.180	-0.192	-0.183	-0.181
		(0.136)	(0.155)	(0.186)	(0.190)	(0.240)	(0.273)
RCT			-0.223*	-0.244*	-0.264*	-0.301**	-0.375
			(0.080)	(0.057)	(0.051)	(0.016)	(0.148)
Matching			-0.242	-0.246	-0.233	-0.165	-0.450
			(0.216)	(0.210)	(0.255)	(0.540)	(0.152)
Primary			-0.218	-0.224	-0.254	-0.254	-0.401
			(0.252)	(0.294)	(0.322)	(0.344)	(0.357)
Work. Paper				0.056	0.034	0.060	0.017
				(0.479)	(0.693)	(0.521)	(0.874)
Report				-0.013	-0.009	0.007	-0.132
				(0.947)	(0.965)	(0.976)	(0.606)
East Afr.					-0.089	-0.080	-0.106
					(0.435)	(0.487)	(0.477)
Southern Afr.					-0.225	-0.234	-0.400
					(0.290)	(0.265)	(0.172)
Natl. Rep.						-0.113	-0.113
						(0.390)	(0.425)
Length (mo.)							-0.004
							(0.413)
Experiments	60	60	60	60	60	59	55
I^2	88.19	85.72	85.87	86.35	86.09	85.72	86.20
t^2	0.033	0.038	0.047	0.054	0.060	0.061	0.068

P(|t|) in parentheses. Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

which children are exposed"; this is in contrast to high income countries where family background variables absorb a large amount of this variation (pg. 1162). Further, it is possible that other interventions in this category such as textbook provision or class size reduction may prove to be relatively less successful if a teacher was not able to adapt his/ her teaching methods in order to take advantage of such interventions. In addition, programs meant to increase learning through increased student and teacher attendance may also be less successful if the quality of instruction itself is not changing. Section C (in this chapter) explores the variation in effect size estimates within pedagogical interventions themselves.

Effectiveness of school health interventions: A large number of studies in recent years have focused on improving student cognitive processing, as well as other health and education outcomes (attendance, enrollment, retention) through the provision of school meals or nutritional supplements, as well as through the provision of health treatments such as preventative malaria treatment, malaria screening and treatment, or the use of anti-helminth (de-worming) drugs. However, interventions in these categories have some of the lowest pooled effect sizes in this dataset and are statistically insignificant (see Table 12 below). Even when pooling these treatments under one "School Health" category, the effect is still low (0.019 standard deviations) and imprecisely measured despite the larger number of studies in this sub-sample. Further, all of the studies in this category are randomized controlled trials, and all rate in the highest 5% of study quality.

Further, when I measure the average impact of school health interventions against all others in the sample, I find that these studies have on average a pooled effect size 0.44 standard deviations lower (SE = 0.197, df = 11.3, p = 0.043) than other interventions, controlling for a host of study-level factors (Table 27, Appendix G). Even when I limit the full sample to high quality studies, the impact is still negative, closer to zero and less precisely measured (Table 28, Appendix G).

Table 12. Pooled effect sizes for school health interventions

	Estimate	Std. Error	df	(P|t|)	95% CI.L	95% CI.U	I^2	t^2	Studies
Student Cognitive Ability (academic outcomes)									
Health Treatment	-0.008	0.041	3.5	‡	-0.13	0.11	45.7	0	5
School Meals & Supplements	0.059	0.022	1.16	‡	-0.15	0.27	0	0	4
School Health†	0.019	0.641	4.67	0.55	-0.06	0.10	27.6	0.002	9

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

†Combined heath treatments & school meals/ supplements.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

However, these effects are all measured using academic outcomes. When I examine the impact of the school health programs on cognitive outcomes such as tests of memory and attention (by limiting the sample to studies that either only examine cognitive outcomes or examine both academic and cognitive outcomes), I find that while there is still not a measurable pooled effect of meals or supplements, health treatments have a reasonably high impact on cognitive outcome (*d* = 0.176, SE = 0.028, df = 2.18) (Table 13 below). Further, it is important to note that the cognitive effects on students with compromised immune systems can be substantially higher, though these effects are not routinely measured. For example, Gee (2010), examines the effect of a malaria treatment program and estimates effect sizes of 0.66 and 0.71 standard deviations on cognitive tests for student infected with the parasite Schistosomiasis.

Table 13. Pooled effect sizes of school health programs on student cognitive performance.

Intervention	Estimate	Std Err	df	P(|t|>)	95% CI.L	95% CI.U	I^2	t^2	Studies
Meals/Supplements	-0.0185	0.012	2.29	‡	-0.07	0.03	0	0	5
Health Treatments	0.176	0.028	2.18	‡	0.07	0.29	0	0	4

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

As to why these impacts are relatively small and imprecise in regards to school meals programs (and in regards to academic outcomes associated with health treatments), there are a variety of hypotheses. Regarding school meals, it is possible that the provision of meals takes time away from instruction (Vermeersch and Kremer, 2004) or that households respond to this provision by re-apportioning food among family members (and possibly away from the child receiving the school lunch) (Jacoby, 2002). It is also possible that as these programs are often associated with increased enrollment and attendance, students not previously attending school may have caused an increase in class size (Vermeersch and Kremer, 2004), or they may have entered with lower initial test scores, thus attenuating these potential effects. Finally, it may be that the assessments used to measure cognition are not sensitive enough to pick up on subtle changes in student cognitive processing (only 50% of these examinations were tested for psychometric reliability). Regarding health treatments, it is possible that prevalence of malaria or helminth infections in these regions was not high enough to cause large changes in student academic achievement and that because these estimates measure intention-to-treat outcomes on all students, the effects found here are not representative of treatment effects on the infected. Finally, given the positive and significant impacts on student achievement for cognitive outcomes, it is possible that these health interventions are truly affecting student cognition but that if the quality of instruction is unchanging, then the effects on academic outcomes may remain relatively low. Section C (below) examines more closely the variation within school meals and health treatment interventions.

C. What Explains Variation in Effectiveness?

The final set of research questions in this dissertation examine heterogeneity both across and within intervention types, in both the full sample and the high-quality sample (quality index score greater than or equal to three). I begin with a brief discussion of full sample heterogeneity. For example, I ask: Do studies which use standardized assessments have a lower pooled effect size measure than those which employ researcher-designed tests? Or do studies that employ randomized controlled trials perform differently on average than those that use quasi-experimental methods? Secondly, I examine effect size variation within each intervention area and attempt to assess why certain interventions appear more successful than others.

1- Analyses of Full Sample Heterogeneity

In this section, I explore determinants of effect size differences in the full sample and the high quality sample. I have included the majority of these variables in meta-regressions above and have found that my results are robust to the inclusion of these variables, but here I highlight heterogeneity statistics (predominately calculated through meta-regressions) to provide a more complete understanding of variation within the dataset as a whole (see Table 29 in Appendix G).

Heterogeneity by methodology: I find that randomized controlled trials in this sample have pooled effect sizes that are on average indistinguishably different from other methodologies used; this holds for both the full sample and the high quality sample. The same holds for studies that employ matching techniques. (Note: RCTs and matching strategies make up 90% of the full sample). Regarding estimators, intention-to-treat estimates are not statistically significantly different from other estimators in this sample (coefficient remains near zero for both the full and high quality sample); this also holds for average treatment on the treated estimators.

Heterogeneity by publication type & journal rank: Compared to studies in peer-reviewed journals, neither working papers nor reports have statistically significantly different pooled effect sizes; In the full sample, the magnitudes of the coefficients suggest that effect sizes in peer reviewed journals may actually be slightly higher, but again these differences are not statistically significant, and they disappear when the sample is restricted to high quality studies. Also, differences in journal rankings do not seem to have a large or statistically significant impact on pooled effect size estimates.

Heterogeneity by assessment: When I compare studies that employ researcher-designed assessments (either purely original or with some questions pulled from standardized tests), I find that the pooled average effect size for these studies is slightly higher than studies that employ standardized tests in the full sample (d = 0.024), but this difference is not significant. And when I limit the sample to only high-quality studies, researcher-designed tests have a pooled effect size 0.05 standard deviations lower than those using standardized test (this difference is also statistically insignificant). One would expect that in this context, standardized test measures would provide less discriminating power than study-administered/ tailored tests (Halpin and Torrente, 2014; Prophet and Badede, 2009; and Hill, 2008), thus it is surprising to see that even in the high-quality sample, there are no significant differences between these groups. This could potentially be due to the fact that so few researcher-designed/adapted tests are examined for reliability (only ~25%), thus contributing to additional noise in the measurement of study constructs.

Further, regarding the psychometric properties of the assessments used in these studies, in the full sample, I find that studies that test for reliability and other psychometric measures have effect sizes over 0.398 standard deviations higher than those that do not (SE = 0.203, df = 16.6 , p = 0.07). However, when the sample is restricted to high quality studies, this difference disappears.

Heterogeneity by location: I find that in the full sample, nationally representative tests have pooled effect sizes 0.177 standard deviations lower than those that are not representative (significant at the 5% level), however this difference again disappears when I limit the sample to only high-quality studies. I also find that in the full sample, studies in both the east and southern African region have lower effect sizes on average than those in west Africa (though these differences are not statistically significant), but the magnitude of these coefficients decreases in the high quality sample, and all differences become statistically insignificant.

Heterogeneity of supply versus demand side interventions: Here I compare the effect sizes of supply-side versus demand-side interventions. For example, are interventions such as school supplies provision, class size reductions, or the hiring of contractual teachers (all supply-side interventions) as effective as cash transfer or scholarship programs (demand-size interventions)? I find that in the full sample, supply side interventions have a much higher pooled effect size (by 0.216 standard deviations) on average than demand-side interventions, though the magnitude of this difference drops somewhat in the high-quality sample (0.065) and is only significant at the 10% level.

Overall, very few moderator variables are statistically significantly correlated with the overall pooled effect size in the high-quality sample. This is particularly surprising in the case of assessment type, as one would expect studies employing standardized test measures to have less discriminating ability and lower effect sizes overall (Halpin and Torrente, 2014; Prophet and Badede, 2009, and Hill, 2008). As this is not the case in either the full sample or the high-quality sample, it is possible that researcher-designed tests are not attaining their full discriminatory potential, possibly due to poor assessment reliability metrics.

2- Heterogeneity within Intervention Type

In this section I explore the results of the above pooled effect sizes and attempt to explain some of the variation in effect size results. Because there are too few studies within many of these intervention areas to be able to run a meta-regression which controls for these moderating variables, I thus combine meta-analytic findings with a supporting narrative review for each intervention area. See forest plots and effect size plots in figures 2-15 (Appendix F) for a visual representation of this data.

a.) Instructional Quality

Pooled effect sizes in this category range from 0.022 standard deviations (SE = 0.015) for school supplies provision to an extremely high pooled effect size of 0.918 standard deviations (SE = 0.314, p = 0.01) for pedagogical interventions, as detailed above (also Table 14).

Table 14. Pooled effects size of instructional interventions

	Estimate	Std. Error	df	(P|t|)	95% CI.L	95% CI.U	I^2	t^2
Quality of Instruction								
School Supplies Provision	0.022	0.015	1.25	‡	-0.10	0.14	0	0
Class Size & Composition	0.109	0.049	1	‡	-0.51	0.73	0	0
Instructional Time†	0.464	0.198	…	…	…	…	…	…
Pedagogical Intervention	0.918**	0.314	15.1	0.01	0.25	1.59	95.6	0.4

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

† There was only one study which focused on the impact of an increase in instructional time. ‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

As pedagogical interventions have the highest effect size and are the most numerous, I begin with a discussion of heterogeneity within these studies.

Pedagogical Interventions: I first attempt to explain some of the variation in effects through sub-group analysis (see Table 15 below). I limit the sample to only high quality studies because using a meta-regression to control for numerous moderating factors is not possible within a topic, due to sample size limitations. When I do this, I find that interventions which employ adaptive instruction (either computer-assisted learning programs which adapt to the student's learning level or pedagogical methods that emphasize formative assessment and instruction that is individualized and targeted), have on average a higher pooled effect size than those interventions that do not employ those techniques; the pooled effect size associated with adaptive instruction is 0.461 standard deviations (SE = 0.162, df = 1.61), while the pooled effect size associated with non-adaptive instruction is 0.123 standard deviations (SE = 0.049, df = 2.31). An example of such a program is the EGRA (full program) in Liberia evaluated by Piper and Korda (2011), in which students' reading levels were evaluated using a diagnostic exam, and teachers were then trained in how to continually assess student progress using the instruments provided (the program also provided pedagogic support, other resource materials, and parents were informed of student achievement levels). Further, both teacher-led and computer-assisted methods have a statistically significant effect (at 10% level) on student performance (0.214 standard deviations, SE = 0.082, df = 4.47, p = 0.054) and 0.436 standard deviations (SE = 0.116, df = 1.69, p= 0.054) standard deviations, respectively).

Among high quality pedagogical interventions that involved teacher training, each of these interventions involved either long-term teacher mentoring or in-school teacher coaching; the pooled effect size of these interventions is 0.249 standard deviations (SE = 0.133, df = 2.68).

This is important to note as “teacher training” is often associated with one-time in-service trainings taking place in a central location and not prolonged one-on-one in-school teacher coaching, as in these programs. An example of a program employing this type of teacher training is the READ program in rural South Africa (evaluated by Sailors et al, 2010) which provides students with high quality books relevant to the students’ lives and provides training to teachers on strategies to integrate these books into their lesson plans; the “intensive and systematic professional development” includes demonstration lessons by READ mentors, monthly coaching visits by READ staff, one-on-one reflections sessions after monitoring visits, and after-school workshops for both teachers and school administrators.

Table 15. Pooled effect sizes for pedagogical sub-samples

Sample = High Quality Studies	Estimate	Std Err	df	P(|t|>)	95% CI.L	95% CI.U	Studies
OVERALL	0.228**	0.078	5.2	0.0317	0.029	0.426	9
Adaptive instruction	0.421	0.162	1.61	‡	-0.467	1.310	3
Non-adaptive	0.123	0.049	2.31	‡	-0.061	0.307	5
Teacher only	0.214*	0.082	4.47	0.054	-0.005	0.434	3
Computer-assisted	0.436	0.116	1.69	‡	-0.163	1.040	6
Coaching or mentoring	0.249	0.133	2.68	‡	-0.202	0.700	4
In-service training	*No data*	...	...	...	...	...	
Class teacher-led	0.216**	0.079	4.82	0.043	0.011	0.421	8
Researcher-led	*No data*	...	...	...	...	...	

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

Finally, to further understand the mechanisms behind the consistently positive impact of pedagogical methods, I examine the pedagogical techniques themselves that may drive the results (Table 16). Unfortunately, a number of these interventions do not detail the pedagogical techniques used in their interventions, nor do they examine the educational theory behind them.

For example, the EGRA program described above (Piper and Korda, 2011) does not provide any further details on the type of "pedagogic support" provided to teachers. When the full sample is limited to studies that do provide this information, I find that inquiry-based learning programs (those that focus on problem-solving skills, are student-centered, and activity-oriented), have an extremely high pooled effect size (3.15 standard deviations, SE = 3.16, df = 1). However, when the sample is limited to only high-quality interventions, only procedural learning interventions (those that break learning down into a set of skills to be mastered sequentially) remain in the sample. These have a pooled effect size of 0.118 standard deviations (SE = 0.043, df = 2.61).

Table 16. Pooled effect sizes for pedagogical sub-samples (instructional techniques)

Pedagogical Style	Estimate	Std Err	df	P(|t|>)	95% CI.L	95% CI.U	I^2	Studies
Bilingual instruction	2.70	2.38	1	‡	-27.5	32.9	2	2
Conceptual learning†	0.461	0.638	...	...	...	...	...	...
Cooperative learning	1.20	0.796	1	‡	-8.91	11.3	2	2
Inquiry-based learning	3.15	3.16	1	‡	-37	43.3	2	2
Procedural learning	0.448	0.265	4.96	0.15	-0.234	1.13	6	6
Procedural learning *(High quality studies only)*	0.118	0.043	2.61	‡	-0.032	0.27	4	4

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

†Only one study examines the impact of conceptual learning.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

Even within high quality procedural learning programs there is a high degree of heterogeneity (I^2 = 59.8). For example, Lucas et al (2013) evaluate the Reading to Learn (RTL) intervention in both Kenya and Uganda. This program uses a "reading scaffolding" technique, a procedural learning approach which is built "from a conceptual understanding of stories, to the decoding of letter-sound relationships, to the eventual written production of new sentences and stories" and is accompanied by teacher mentoring and literacy material provision. Though the program design was equivalent, the impact of the program ranged from 0.08 standard deviations

(SE = 0.064) in reading in Kenya to 0.20 standard deviations (SE = 0.054) in Uganda. In this case, the authors attribute this differential not to contextual differences (such as beginning literacy levels between countries) or differences in degrees of program implementation, but to the fact that in Kenya, students were testing in Swahili, which is the official but not the de facto language of instruction.

While exploring the heterogeneity within this topic is limited both by quality issues and sample size, there is suggestive evidence that programs that employ adaptive learning techniques and teacher training programs that feature one-on-one mentoring can have consistently positive effects. Finally, the average length of interventions in this area is approximately 7-8 months, considerably shorter than interventions in other areas which are on average 1.5 years.

Instructional Time: Only one study resulting from my systematic search examined the impact of instructional time on student performance. Orkin (2013) takes advantage of a natural experiment in Ethiopia which increased the daily length of instructional time by 30% (approximately 2 hours) to perform a difference-in-difference analysis of this policy change (there was large variation in the number of schools implementing this change over time). After converting her results to the *d*-family effect size unit (standardized mean differences), we see that the instructional time increase improved student test scores in Mathematics (*d* = 0.412 standard deviations, SE = 0.184), Reading (*d* = 0.119 standard deviations, SE = 0.20) and Writing (*d* = 0.861 standard deviations, SE = 0.201), with a composite effect of 0.464 standard deviations (SE = 0.198), which is a large effect size for this literature. However, the assessment used to measure student achievement in this study consisted of only a handful of questions testing numeracy and reading and was not subject to any psychometric testing. And while district discretion or impetus to move forward with the reform generated variability in reform implementation at the school level, it also raises questions regarding true exogenous variation in policy implementation. For these quality-related reasons, this result should be interpreted cautiously.

Class Size & Composition: Experimental studies examining the impact of class size reductions and student tracking/ peer effects on performance are also rare in Sub-Saharan Africa. The pooled effect size for these interventions is 0.109 standard deviations (SE = 0.049, df = 1). In related but independent randomized experiments, Duflo, Dupas and Kremer (2011) examine the impact of student tracking through ability grouping, while Duflo, Dupas and Kremer (2012) examine the impact of class size on student performance (among other interventions). Duflo, Dupas and Kremer (2011) find that tracking students into two performance groups results in an average impact of 0.166 standard deviations (SE = 0.098) in language and 0.156 standard deviations (SE = 0.083) in math. They find that even for low-performing students, tracking improved student performance by 0.156 standard deviations (SE = 0.075) and that these results carried over into the next school year after the program had stopped. They hypothesize that these results are due to the fact that teachers are better able to adapt their teaching methods to more homogeneous student populations. However, a paper uncovered in this search that was not included in this meta-analysis (given that it was the only tertiary-level study found), examines the impact of a policy that “de-tracked” students out of ability-grouped living quarters into mixed ability dormitories and found large performance effects among all students (0.123 standard deviations, SE = 0.031), with even larger effects on low-performing students (0.224 standard deviations, SE = 0.075); the author hypothesizes that these effects are due to peer-to-peer interactions with students of different levels and studying habits (Garlick, 2013). Thus, while peer effects may have some positive impact on students, as suggested by Garlick (2013), these effects may be trumped in the case of Duflo, Dupas and Kremer (2011) by teacher preferences for homogeneous student learning levels.

Regarding class size, Duflo, Dupas and Kremer (2012) examine the impact of a class size reduction from 82 to 44, and find a relatively low and statistically insignificant impact of such an intervention (0.051 standard deviations [SE = 0.078] in arithmetic and 0.102 standard deviations [SE = 0.111] in language). However, this effect is somewhat confounded by the fact that this class size experiment was run in parallel with a teacher contract experiment; the class size effect the authors are measuring is the effect of a reduction in class size (in classrooms with civil service teachers) in schools that received a contract teacher. Because these civil service teachers "reduced their effort in response to the drop in the pupil-teacher ratio," this class size effect may be under-estimated (abstract). The authors do try to tease out these effects using instrumental variables methods and estimate that the true impact of the class size reduction is closer to 0.042 - 0.064 standard deviations (statistically significant at the 5% level) for a 10-student reduction (and thus 0.168 - 0.256 for a 40 student reduction, assuming linear effects), however the authors themselves question whether the exclusion restrictions hold in these estimations, saying that "none of these exclusion restrictions are perfect, so the exercise is more illustrative than absolutely definitive," thus the true impact of class size reduction is yet to be clearly examined in any context in Sub-Saharan Africa (pg. 17).

School Supplies: School supply interventions range from the provision of textbooks or flipcharts to school grants that are ear-marked for learning materials only (notebooks, textbooks etc.). The pooled effect size for these interventions (measured on all students) is 0.022 standard deviations (SE = 0.015, df = 1.25). These results are on average lower than a number of the other program types in this study, but the effects vary quite a bit within this category. For example, a randomized experiment of textbook provision in Kenya seemed to only have an impact when given to students performing in the highest quintile (impact of 0.218 standard deviations in year one (SE = 0.096) and an impact of 0.173 standard deviations (SE = 0.131) in year two) (Glewwe, Kremer, and Moulin, 2009). Meanwhile, a non-parametric matching study across five francophone countries found that the impact of textbooks on student language performance was quite high yet statistically insignificant (0.264 standard deviations, SE = 0.283), with a particularly high estimate for the impact of textbooks on students in rural areas (0.456 standard deviations, SE = 0.346) (Frölich & Michaelowa, 2011). These differences may stem from different starting literacy levels of the student populations, varying levels of difficulty of the textbooks, or differences in how teachers integrated these books into the curriculum (among others reasons), but it is clear that textbooks distribution may affect different sub-populations in ways that differ greatly from the average pooled effect.

In addition, flipchart provision to schools (through a randomized experiment) (Glewwe et al, 2004) also had a very low (0.008 standard deviations) and insignificant average impact on student learning, possibly due to the fact that they were not used intensively for all of the grades or subjects for which they were intended. Finally, using a natural experiment, Das et al (2013) examine the impact of anticipated and unanticipated school grants (to be used for textbooks and other schools supplies) in Zambia and find that unanticipated school grants (approx. $3 per student) have a larger effect on student performance (0.10 standard deviations in both math (SE = 0.048) and language (SE = 0.050)) than those that are anticipated. The authors argue that these findings are due to the behavioral response of households who decrease education spending in response to anticipated grants. They suggest that the policy implication of this result is to invest in inputs whose impact may be less likely to be attenuated by household substitution such as teacher inputs (teacher training, use of contract teachers) and improving classroom pedagogy.

b. Cognitive Processing Abilities

As fully explored in Section B, Part C "Robustness tests of high and low impact intervention types" above, interventions in the categories of health treatments and school meals had some of the lowest pooled effect sizes (for academic outcomes) in this dataset and were statistically insignificant (see Table 17 below). However, I also found that when the sample is limited to studies that examined at least one cognitive outcome, the pooled effect size of health treatments does rise (see Table 18 below). In this section, I focus on heterogeneity across and within the categories in this group.

Table 17. Pooled effect sizes for school health treatments

	Estimate	Std. Error	df	(P|t|)	95% CI.L	95% CI.U	I^2	t^2	Studies
Student Cognitive Ability (academic outcomes)									
Health Treatment	-0.008	0.041	3.5	‡	-0.13	0.11	45.7	0	5
School Meals & Supplements	0.059	0.022	1.16	‡	-0.15	0.27	0	0	4
School Health†	0.019	0.641	4.67	0.55	-0.06	0.10	27.6	0.002	9

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

†Combined heath treatments & school meals/ supplements. ‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and values should not be interpreted.

Table 18. Pooled effects of school health programs on student cognitive performance

Intervention	Estimate	Std Err	df	P(|t|>)	95% CI.L	95% CI.U	I^2	t^2	Studies
Meals/Supplements	-0.0185	0.012	2.29	‡	-0.07	0.03	0	0	5
Health Treatments	0.176	0.028	2.18	‡	0.07	0.29	0	0	4

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

First, when I examine these effects comparatively, I find that school meals and nutritional supplement programs have a slightly higher pooled effect size than health treatments overall (0.051 standard deviations), but this difference is not significant (SE = 0.053, df = 3.69) (Table 19 below).

Table 19. Differential effect of school meals programs (versus health treatments)

	Estimate	StdErr	df	P(|t|>)	95% CI.L	95% CI.U
Intercept	-0.0048	0.044	3.2	‡	-0.14	0.13
School Meals & Supplements	0.0512	0.053	3.69	‡	-0.102	0.204

Number of studies = 9

Number of outcomes = 23 (min = 1 , mean = 2.56 , median = 2 , max = 6)

ρ = 0.8

I^2 = 23.015

τ^2 estimate = 0.002

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

School meals and supplements: Regarding school meals and supplements, there was some variability in the effect sizes of these interventions. Those programs that provided school breakfast, lunch, or take-home rations generally had effect sizes that ranged from 0.02 on a written curricular test (not significant) (Vermeersch and Kremer, 2004) for a school breakfast program to 0.09 standard deviations (significant at the 1% level) (Kazianga, deWalque, and Alderman, 2012) for a school lunch program. Vermeersch and Kremer (2004) do find, however, that the impact of the school meals program was higher (0.41 standard deviations) in schools with teachers who had more experience, possibly suggesting that time management skills of experienced teachers might have facilitated the implementation of the school meals program. Exceptions to these relatively low average effect size estimates include Whaley et al (2003) who found that mid-morning supplements (particularly beans with either meat or oil) had high magnitude effect sizes (up to 0.18 in math for the beans and meat supplement and 0.25 standard deviations in math for the beans and oil diet); however, these effect sizes are highly imprecise (see the effect size plot in Figure 7, Appendix F). Meanwhile programs that tried to isolate the impact of additional nutritional supplements and vitamins (by randomizing students between fortified and non-fortified biscuits) had effect sizes that were between -0.12 to 0.03 standard deviations (all statistically insignificant), even for cases of iron-deficient students (Baumgartner et al, 2012). Further, when I limit the sample to those studies that focus on cognitive outcomes, I find that there is also no differential effect of program length on program effectiveness (Table 30, Appendix G).

Health treatments: Examining the variation in effect sizes within health treatments (using cognitive outcomes), I find that interventions that target malaria are driving these results, as there is only one helminth treatment intervention that evaluates its impact on cognitive outcomes. When I calculate the pooled effect size for the malaria interventions on cognitive outcomes, I find an average impact of 0.189 standard deviations (SE = 0.022, df = 1.57) (Table 20 below).

Table 20. Pooled effect of malaria treatment on cognitive processing

Sample= Health treatments only	Estimate	StdErr	t-val	df	P(|t|>)	95% CI.L	95% CI.U
Malaria treatment or prophylaxis	0.189	0.022	8.47	1.57	‡	0.063	0.314

Number of studies = 3
Number of outcomes = 5 (min = 1 , mean = 1.67 , median = 2 , max = 2)
$\rho = 0.8$
$I^2 = 0$
τ^2 estimate = 0

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

c. Student and Teacher Motivation

The question underlying this group of studies is whether increased student or teacher motivation through performance incentives can improve student learning. The pooled effect size of these interventions is 0.288 (SE = 0.015, df = 1) for student incentives and 0.075 (SE = 0.047, df = 4.89, p = 0.2) for teacher incentives, though there is a lot of heterogeneity within teacher incentives interventions (see Table 21 below). On average, student incentives have a higher pooled effect size (than teacher incentives) by 0.210 standard deviations (SE = 0.050), however, there are very few interventions in the student incentives category (see Table 22 below). Below I thus provide narrative reviews of these topics, which attempt to explain some of this variation.

Table 21. Pooled effect sizes of student and teacher incentive programs

		Std.			95%	95%		
	Estimate	Error	df	(P|t|)	CI.L	CI.U	I^2	t^2
Student/ Teacher Motivation								
Student Incentives	0.288	0.015	1	‡	0.10	0.48	0	0
Teacher Incentives	0.075	0.047	4.89	0.2	-0.05	0.20	53.45	0

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

Table 22. Differential impact of student incentives (versus teacher incentives)

Sample=						95%	95%
Incentive studies only	Estimate	std. err.	t-val	df	P(|t|>)	CI.L	CI.U
Intercept	0.077	0.047	1.62	4.79	0.17	-0.05	0.2
Student Incentives	0.210	0.050	4.23	1.42	‡	-0.11	0.534

Number of studies = 10

Number of outcomes = 23 (min = 1 , mean = 2.3 , median = 2 , max = 4)

$\rho = 0.8$

$I^2 = 46.55$

τ^2 estimate = 0.010

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

Student Incentives: On average, student performance incentives had a pooled effect size of 0.288 standard deviations (SE = 0.015, df = 1), which is one of the highest pooled effect sizes in the sample. While this estimate is significant at the 5% level, there are only two studies examining the impact of student incentives on achievement (Blimpo, 2010 and Kremer, Miguel and Thornton, 2009) and thus the degrees of freedom resulting from this RVE estimate limits its interpretation. Examining the plot of effect size estimates of student incentives interventions (see Figure 9, Appendix F), I see that the estimates across all student incentives types are very comparable. Regardless of the type of incentive (various student monetary incentives in Benin or a girls merit scholarship competition in Kenya), the effect size estimates range from between 0.27 (SE = 0.16) for girls in Busia district in Kenya,[8] to 0.34 standard deviations (SE = 0.13) for students in the team tournament group in Benin. Further, a range of study characteristics are comparable between these two papers: upper primary school students were the population of interest in both studies, both studies used a RCT methodology, and both interventions lasted approximately 12 months.

Due to the use of multiple treatment arms, Blimpo (2010) is able to compare the relative effectiveness of different types of incentives: monetary incentives for individual students who meet a performance target (US$10 with the potential for a US$30 bonus), team-based incentives in which all students within a team must meet a performance target (US$40 with the potential for a US$120 bonus), or team-based tournaments in which students are competing against other teams for a performance incentive (US$640). Blimpo found that team tournament incentives had the highest impact on student achievement, due to potentially maximized peer-to-peer tutoring within teams (0.34 standard deviations, SE = 0.13), but even the individual incentive schemes

[8] These are the results for the ITT sample. I cite the results from Busia district here. Note that results for Teso district were invalid given that multiple schools withdrew from the program; authors cannot reject the hypothesis of no program effect in Teso.

were quite successful (impact of 0.29 standard deviations, SE = 0.12). The merit scholarship in Kenya was worth the equivalent of US$6.40 per student paid to the school with US$12.80 per student for school supplies paid to each family (plus public recognition). Thus, both programs show that monetary incentives for the individual (US$10/ student) or school fee-related payments of approximately US$20 per student (total) can motivate students to perform at higher levels.

Teacher Incentives: There have been a number of interventions conducted in the area of teacher incentives; these include studies of short-term teacher contracts (i.e, fixed term contracts with lower remuneration, reduced entry requirements, and often with potential for renewal), teacher pay for performance-based incentives, and increases in teacher salaries. The pooled average effect size of these interventions is estimated to be 0.075 standard deviations (SE = 0.047, df = 4.89, p = 0.2), with a high degree of heterogeneity (I^2 = 53.45).

Results from short-term teacher contract studies are varied. Those with the largest and statically significant results are from a randomized trial in Kenya (Duflo, Dupas, and Kremer, 2012) where teacher contracts (2-years, non-renewable, US$120/ month) increased student performance in math (by 0.26 standard deviations) and language (by 0.18 standard deviations), both significant at the 5% and 1% level, respectively. Bold et al (2013a) also find that a scaled-up version of this same contract teacher program in Kenya resulted in an increase of 0.175 standard deviations (SE = 0.091) in student performance (composite score) when implemented by non-governmental organizations. However, they found that when this same program was implemented by government agencies at a larger scale, this effect fell to -0.02 standard deviations (SE = 0.095), suggesting that this difference is due to weak public implementation.

Bourdon, Frölich and Michaelowa (2010) also suggest that the varying impact of short-term teacher contracts may be linked to implementation issues. Using non-parametric matching and PASEC (West Africa standardized tests) data, they find that the impact of similar teacher contract programs (with comparable wages) varies considerably from country to country. They find mixed effects in Togo, largely negative effects in Niger, and large, positive but very noisy effects in Mali (0.710 standard deviations is the average composite measure with a standard error of 0.745). The authors hypothesize that these differences can largely be explained by the fact that contract teachers in Mali and Togo were introduced in a decentralized system which gave more oversight to the local school community in terms of monitoring and hiring responsibilities, whereas in Niger the centralized approach may have resulted in lower teacher accountability and motivation.

Regarding performance-based incentives, Glewwe, Ilias, Kremer (2010) find that incentives that are linked to student performance on a particular examination can effectively increase performance on that particular test (impact estimated to be 0.14 standard deviations, significant at the 10% level), but when the same set of students takes a test which is not directly linked to these incentives, they find that this effect becomes close to zero and insignificant. They find that the effect on examinations linked to incentives is partially due to increased teacher test preparation sessions. However, the reliability of the test not linked to incentives (designed by an NGO) is not reported here, but if this test produced noisy measures, this could also explain the imprecise and null effects associated with the non-linked estimate.

Finally, differences in teacher salaries is also explored by Bold et al (2013a); they find that the impact of earning US$121 versus US$67 (the average income of a teacher hired by the school management committee), has no measurable effect on performance levels of short-term contract teachers. In all, these studies show that short-term teacher contracts can be effective under the right set of local conditions and that while performance incentives have been shown to have positive impacts, these programs should be cautious of the ways in which teachers may feel pressured to teach a more narrow and tailored set of skills to students.

d. Student or Community Financial Restrictions

Interventions that reduce the financial burden on students, their families, and the community can include cash transfer programs (in which the transfers may be either conditional on school attendance or completely unconditional), school uniform provision (uniforms are a required expense in many countries), the abolishment of tuition-related school fees by the government, or the provision of a school (with other complementary inputs) in areas that are under-resourced and under-served.

Table 23. Pooled effect sizes of cost reduction and infrastructure programs

	Estimate	Std. Error	df	(P|t|)	95% CI.L	95% CI.U	I^2	t^2
Cost Reduction Intervention	0.036	0.036	1.58	‡	-0.16	0.24	27.3	0
Infrastructure + Add. Inputs	0.189	0.122	1.97	‡	-0.35	0.72	94.23	0.1

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

Cost Reduction Interventions: Interventions that reduce cost directly to students and families through cash transfers, school fee elimination, or uniform provision have a modest pooled effect size of 0.036 standard deviations (SE = 0.036), and there is relatively low heterogeneity within this group of studies (I^2 = 27.3). Regarding cash transfer programs, in a randomized experiment Baird, McIntosh, and Ozler (2011) find that a conditional cash transfer program (US$4-10 per month) for girls in Malawi had on average higher and more precisely estimated effects on student learning (and drop-out rates) than an identical unconditional program, particularly in the case of language achievement (effect sizes linked to the conditional program ranged from 0.09 to 0.12 for academic outcomes, versus -0.03 to 0.06 for the unconditional programs). However, it is important to note that the authors do find substantial positive effects of these programs (particularly unconditional transfers) on teenage pregnancy and marriage. Further, a propensity score matching study of unconditional support grants ($30/month) in South Africa found quite a high impact on girls' performance (0.23 standard deviations), with a lower impact on boys (0.09), though neither of these estimates was statistically significant at the 5% level.[9]

Meanwhile, Lucas and Mbiti (2012) found that the implementation of free primary education had no effect on student performance, which they say is actually a positive result, given that this policy also increased enrollment, particularly for students from disadvantaged backgrounds which may have also worked to attenuate the performance impact (as was the case with a large Cambodian scholarship program (Filmer and Schady, 2009)). The largest magnitude impact was measured by Evans, Kremer, and Ngatia (2009) for the impact of school uniforms on student performance (0.25 standard deviations, significant at the 5% level), an impact which persisted into year two (0.18), though did not continue to be statistically significant. This large effect is hypothesized to be due to the increased attendance rates of students who received a uniform (attendance rates for these students increased by 44% for the average student), though the authors cannot rule out the possibility that some of this effect could be due to the psychological impact of “winning” a uniform, as winners were termed “sponsor children” by the NGO implementing this program.

[9] Note: this is the impact for children enrolled at birth ("treatment”), versus those enrolled at age six ("control"); Table 7.15 (pg. 60); Uses local linear matching.

Infrastructure with complementary inputs: While the overall effect size estimate for this group of studies is high, it is also highly heterogeneous ($I^2 = 94.23$). This may be due to the fact that although these interventions were all predominately infrastructure interventions, they were accompanied by different sets of complimentary interventions, whose implementation fidelity also varied from site to site. For example, Kazianga et al (2013) and Dumitrescu et al (2011) both evaluate primary school construction programs aimed to increase girls' educational outcomes in Burkina Faso and Niger, respectively. In Niger, the infrastructure project in each village included: three classrooms, housing for three female teachers, a preschool, and separate latrines for boys and girls. Complementary inputs included training for school management committees, local officials, and teachers (with modules), as well as the promotion of extra-curricular activities, and a campaign to raise awareness about the importance of girls' education. However, these complementary interventions were not fully implemented due to political issues. In Burkina Faso, the intervention consisted of similar infrastructure projects, plus school meals & rations, school resources (textbooks), mobilization campaigns, community monitoring, adult literacy programs, and female mentoring (and were implemented with reasonable fidelity).

While the program in Niger had relatively low and imprecisely measured effects (ranging from 0.05 in Math to 0.09 in Language), the program in Burkina Faso had a very large impact on student achievement (0.41 standard deviations). Dumitrescu et al (2011) hypothesize that this difference could have been due to different initial conditions within rural Burkina Faso and Niger (the authors state that the schools constructed in villages in Niger were often in addition to schools already present and that communities had not requested these schools, as was the case in Burkina Faso). Also, the fact that the complementary interventions were not fully implemented in Niger may have also had a large impact on the relatively low levels of performance progress.

Thus it is difficult to tease out whether these differences arise from variation in implementation fidelity or initial conditions. Finally, Martinez, Naudeau, and Pereira (2012) evaluate a pre-school building program in Mozambique. Complementary interventions here also include two volunteer teachers selected by the school management committee, training to parents and teachers, and other community support. The effect of this program on academic outcomes (language) was relatively low (0.11) and imprecisely estimated, though the impact on cognitive scores was quite high (0.25 standard deviations) and significant at traditional levels.

In all, programs that have attempted to raise student achievement through cash transfers have been moderately successful on average, and there is limited evidence that suggests that conditional cash transfers may target educational outcomes over unconditional transfers. Further provision of school uniforms is shown to be a particularly effective way to increase student attendance and thereby student achievement. Finally, programs that are predominately infrastructure-based but that offer a whole host of complementary inputs have been shown to have the potential for very large effects when they are well targeted and fully implemented.

e. School or System Accountability

Interventions that try to increase school and system-level accountability include both information-related interventions (information on school performance or funding provided to the community), as well as interventions that involve school-based or district-based management.

Table 24. Pooled effect sizes of management & information provision programs

	Estimate	Std. Error	df	(P|t|)	95% CI.L	95% CI.U	I^2	t^2
School/ System Accountability								
Management Intervention	0.016	0.028	3.14	‡	-0.71	0.10	18.72	0
Information Provision	0.147	0.053	1.63	‡	-0.14	0.43	0	0

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

Information for Accountability: The average pooled effect of these information provision intervention is relatively large at 0.147 standard deviations (SE = 0.053, df = 1.63) (see Table 24 above). However, the kinds of information interventions within this category vary considerably. Both Reinikka and Svensson (2011) and Bjorkman (2006) examine a government program in Uganda (through difference-in-difference and instrumental variables methodologies) in which a newspaper campaign was launched in order to bring attention to the amount of funding that local schools should receive, aiming to increase transparency and reduce capture. Both studies find that the provision of such information had a large impact on student performance (estimates range between approximately 0.33 standard deviation from Bjorkman (2006) to 0.58 standard deviations for Reinikka and Svensson), though in the case of Reinikka and Svensson (2011), this estimate is not statistically significant. However, Hubbard (2007) questions whether the entirety of the impacts estimated in these studies can be attributed to the newspaper campaign alone, as the information campaign was part of a broader education and fiscal reform strategy (including universal primary education).

Nguyen (2008) evaluates a very different type of information campaign in Madagascar, in which students are informed of the rates of return to education, in an effort to update students' perceived returns to education, increase student attendance, and improve performance. He finds that providing (through class teachers) simple statistics regarding the monetary returns to education have a large and statistically significant impact on student achievement (estimates range from 0.24-0.26). He finds that when this information is shared by an individual that the student could consider as a "role model," the effects are lower (hover around 0.10), except in the case of a role model from a low-income background sharing this information with a student from a similar background. Finally, Piper and Korda (2011) conduct an information intervention in

Liberia (as part of a larger pedagogical intervention) in which student and school report cards are made available to the community, in an effort to hold teachers and schools accountable for student reading progress, and they find relatively modest effects (0.04 standard deviations). From this limited sample of studies, it does appear that the provision of school funding information to communities as well as the provision of rates of return information to students themselves may be promising interventions.

Management interventions: All of the management evaluations in this sample are randomized controlled trials that take place over a period of 12 months or longer. The pooled effect size for this group of studies is relatively modest at 0.016 (SE = 0.028) with relatively low heterogeneity (I^2 = 18.7%). Programs with relatively low effects include studies of district and school-based management in Madagascar. Lassibille et al (2010) find that programs that provide training, as well as provide tools, guidebooks and report cards to schools, sub-districts, and districts have a relatively modest effect on student performance (0.01-0.05 standard deviations), while interventions that provide management support to districts and sub-districts alone have an even lower estimated magnitude (-0.02 to 0.01 standard deviations), none of which are precisely estimated. Further, Glewwe and Maiga (2011) find similar modest and statistically insignificant effects when examining these management reforms in Madagascar, regardless of the type of teacher hired (civil service or contract).

Programs that provide school committees with real responsibilities and ownership over their schools have more varied effects. Bold et al (2013a), who examine a program in which SMCs were given responsibility to recruit and pay contract teachers and Blimpo and Evans (2011), who examine a program in which SMCs are given grants to be used for school improvement, find negative and imprecise estimates ranging from -0.04 to -0.12. However, Barr et al (2012) find positive results for programs that engage SMCs in the creation of a school monitoring plan (impacts range from 0.11 for monitoring plans that did not involve participation to 0.22 for plans that were fully participatory). In addition, Duflo, Dupas, and Kremer (2012) also found positive results for a school-based management program in which SMCs were given hiring responsibilities; however these results were only measured in schools that were part of a teacher contract program and are imprecise. Among other explanations for these discrepancies is varying degrees of community capacity. Blimpo and Evans (2011) shed some light on this issue with their findings that the effects of the school-based management program in The Gambia were highly moderated by adult literacy rates, suggesting that such programs could contribute to either an increase or decrease in student learning, depending on community capacity.

f. Private School Advantage

Studies of "private school" advantage are not included in the full sample of this meta-analysis, as they do not represent actionable "interventions" or policy changes, but are measures of the performance of private schools in comparison to public or government schools in a state of general equilibrium. However, I do provide a brief summary of these results here. Only two studies of private schools were found in my systematic search, one in Kenya and the other in Nigeria. The authors attempt to get around selection bias issues associated with this kind of analysis by using quasi-experimental techniques (a time series model in Kenya and propensity score matching techniques in Nigeria). The authors find that private schools in both countries have a large advantage and significant effect over government schools – as high as 0.98 standard deviations (on a composite test) in the case of Kenya (Bold et al, 2013b) and as high as 1.13 standard deviations (language results) in Nigeria (Tooley et al, 2011). Tooley et al also find that registered low-fee private schools do not consistently perform better than their unregistered counterparts. Thus the potential performance differential between public and private schools is estimated to be quite high, but it is unclear if this degree of private school advantage would be equivalent in other countries.

D. Publication Bias

In order to detect publication bias in this group of studies in their entirety, I employ funnel plots with associated Egger tests. For this analysis I do not use robust variance estimation, as this type of analysis is not yet available in any statistical software package for publication bias. Funnel plots provide a visual diagnostic of publication bias by showing gaps in the number of small studies with small effects (or effects to the left of the average pooled effect and with high standard errors), which are less likely to be published. The Egger test detects asymmetry in the funnel plot by examining a regression of the standardized effect estimates against their precision and determining whether the intercept deviates significantly from zero.

Funnel plot & Egger tests: While roughly half of the studies included in this meta-analysis are published in journals, I check first check for publication bias within this group of studies as a whole. When I run a funnel plot on the full sample, I notice that small studies (those with large standard errors) are largely absent from the left side of the average pooled effect size (see the plot series in Figure 16, Appendix H), indicative of publication bias. Using an Egger's test, I do find evidence of significant bias, both in the full and high quality sample, though the bias is less in the high quality sample (coefficient on "bias" is 1.84 in full sample, significant at the 1% level versus 0.795 in high quality sample, also significant at the 1% level). When I limit the sample to only those studies that have been published in journals, the bias is still quite high in the full sample but drops in the high quality sample, though it is still significant (see the plot series in Figure 17, Appendix H).

Further, when I examine the extent of publication bias by academic field, I find similar patterns across the economics, education, and public health literature, both in the full sample and high quality sample (see the plot series and Egger tests in Figure 18, Appendix H). However, publication bias is more prevalent in the economics and public health literature than in the education literature, as the bias measured in the education literature is not significant ($p > 0.30$), and the coefficient on this bias is lower in the high quality sample. Further, when I examine publication bias by topic, intervention area, or by methodology type (see the plot series in Figures 19 and 20 (Appendix H), respectively), I again find similar patterns of publication bias across almost all cuts of the data. An exception to this trend is the student incentives literature; however, there are too few studies within this literature to either confirm or reject publication bias.

Fail-safe N: Following Orwin (1983), I calculate a "fail-safe N" and estimate that I would need to find 302 studies with no effect at all in order for the average effect size (0.181 standard deviations) to be driven down to 0.03 standard deviations, which education economists in this field consider practically insignificant.[10] The calculation of a "fail-safe N," generally assumes no large negative effects, which seems reasonable given the distribution of effect sizes in this data (while there are a few negative findings, they are largely insignificant). Further, for pedagogical studies in particular, I would need to find 503 studies with no effect in order for the average effect size (0.918) to be driven down to 0.03.[11] The "fail-safe N" for high quality pedagogical studies is also relatively high at 59.[12] Thus in this sense, publication bias may not be

[10] [Fail safe N for full sample = 60 [(0.181/0.03)-1] = 302].

[11] [Fail safe N for pedagogical studies (all) = 17 [(0.918/0.03)-1] = 503].

[12] [Fail safe N for pedagogical studies (high-quality) = 9 [(0.228/0.03)-1] = 59].

a large cause for concern at this point, however the pooled effect sizes calculated in this meta-analysis may be slightly over-stated.

CHAPTER VI.

DISCUSSION

The focus of this dissertation has been on understanding the body of experimental and quasi-experimental research on learning outcomes in Sub-Saharan Africa. In this section, I discuss the main results of this dissertation, compare the findings of this meta-analysis to previous syntheses, evaluate the limitations of this research, and provide recommendations for future practice and research.

A. Summary of findings

This summary discusses each of this paper's main research questions in turn and then briefly summarizes the findings regarding publication bias.

Research question A: What is the state of the literature in this field? I find that while there is more rigorous evidence on interventions in education than previously reported (particularly in the area of pedagogical methods), this evidence is not evenly distributed across topics or countries. In fact, certain topics that are very relevant to the African context (such as multi-shift teaching, multi-grade teaching, high stakes testing, language of instruction, and class size) are not under rigorous study at all (there is no more than one study in each of these areas). It is unclear why such a dearth of evidence exists in these areas. It is possible that the current research agenda within developed countries may have shaped the direction of recent research in Sub-Saharan Africa, but this is only a hypothesis. Further, much of this research comes from a sample of six countries: Kenya, Nigeria, South Africa, Uganda, Burkina Faso, and Madagascar. These countries are relatively well dispersed throughout Sub-Saharan Africa but are still not representative of the student population in other less-researched countries, particularly those that are war-torn.

In addition, there are numerous descriptive differences in study characteristics across academic fields (i.e. economics, public health, and education) (see Chapter V, Section A, Part 2 for more details). For example, studies from the field of economics (versus public health and education) are more likely to use “composite” measures of performance (combination of math and language scores), are slightly more likely to be nationally representative, are more likely to be working papers than peer-reviewed journal articles, and are more likely to use quasi-experimental methods. Further, 80% of studies from the field of education examine the psychometrics of the assessments used, versus only 2.4% of studies in economics and 50% in public health. Studies in the field of education are also smaller on average (in terms of randomized units) and shorter in length (average length of interventions is 7-8 months).

Research question B: What is the relative effectiveness of intervention types in this sample? To begin, I find that interventions in this sample have a pooled effect size of 0.181 standard deviations (SE = 0.045, df = 47.7, p = 0.0002). In Cohen’s terms, this is considered a “small” effect, though it is close to the class-size effect (0.20) found in the Tennessee class-size study which is often used as a benchmark in education experiments. Importantly, there is a large amount of heterogeneity in true effects across studies (I^2 = 89.32). Then, looking across intervention areas, I identify groups of studies with both extremely high effect sizes: studies in pedagogical methods, and those with seemingly low pooled effect sizes: studies in school health (see Chapter V, Section B, Table 8 for full set of results by intervention area).

First, I find that shifts in **pedagogical methods** have been shown to have a strong and robust effect on student performance. In the full sample, the pooled effect size estimate is extremely large at 0.918 standard deviations (SE = 0.314, df= 15.1, p = 0.01), though when I restrict the sample to only high quality studies, this estimate drops to 0.228 standard deviations (SE = 0.078, df= 5.2, p = 0.032), which is still a large impact in comparison to other intervention types in the sample. Further, when I estimate the average differential impact of pedagogical interventions (using meta-regression in the full sample), I find that these interventions have an effect size approximately 0.30 standard deviations higher than all other interventions in the sample; this result is robust to a number of study and intervention-specific moderator variables, including study quality. The high pooled effect size findings associated with these pedagogical interventions may be surprising - but also intuitive, given the low levels of performance in much of Sub-Saharan Africa (detailed in Chapter II) and the poor state of teacher pedagogy in many schools.

Regarding the low pooled effect sizes associated with **school health programs** (health treatments and school meals), I first estimate the pooled effect size for health interventions alone to be -0.008 standard deviations (SE = 0.041, df = 3.5). When I combine health treatments with school meals, I find the pooled effect size is 0.019 standard deviations, which is still not statistically significant (SE = 0.641, df = 4.67, p = 0.55). And if I run a meta-regression in the high quality sample with school health interventions as my independent variable of interest, I find that the pooled effect size of these interventions is on average 0.119 standard deviations lower than all others in the sample (SE = 0.057, df = 10, p = 0.064).

Regarding health treatment programs, the low magnitude pooled effect size estimate may be partially due to the fact that these estimates are intention-to-treat estimates and not "treatment on the infected" estimates. However, when the sample is restricted to studies with cognitive outcomes, I find that health treatments have a reasonable impact on student cognitive performance (tests of memory and attention), suggesting that these interventions are working to improve cognition, but that if instruction itself is unchanged, academic outcomes will not improve. Concerning school meals, the average pooled effect size for these programs is modest (0.059 standard deviations) and statistically insignificant, even when examining only cognitive outcomes. Possible explanations for the lack of strong effects overall are detailed in Section B, Part 3 (of Chapter V) but include the fact that these school meals programs may have decreased instructional time, increased enrollment, or caused households to reapportion caloric intake.

Research questions C: What explains variation in study effectiveness? I first examine moderators that may explain full sample heterogeneity and then examine variation within each topic. To begin, when I examine heterogeneity statistics within only the high quality sample, I find that study methodology, impact estimators (intention-to-treat, average treatment on the treated etc.), publication type, and region of study are not statistically significantly correlated with pooled effect size measures, nor are the magnitudes of these coefficients large. In addition, assessment type is similarly not correlated with pooled effect size measures, which is surprising given that one would expect that tests tailored to a particular intervention would show more discriminating power than standardized test measures (Halpin and Torrente, 2014 and Prophet and Badede, 2009), suggesting that these researcher-created tests may not be optimally written.

In order to address and understand heterogeneity within each intervention group, I use meta-analysis methods when there are enough studies to do so (e.g., pedagogical methods and school health), and in other areas I turn to narrative reviews, particularly in the case of intervention areas with very few studies. I examine all effects by the "learning channel" through which these interventions are designed to affect student performance (see “Theory of Change”, Chapter III).

Instructional quality: Within studies on instructional quality, I first examine heterogeneity across studies of **pedagogical methods**. I find that studies which employ adaptive instruction and teacher long-term mentoring or coaching may be driving these results, though the sample sizes for these additional analyses is quite small (9 studies). In addition, I find that both teacher-led pedagogical changes and blended (or technology-assisted) learning methods have large pooled effect size estimates and that inquiry-based learning techniques report very large effects in the full sample. However, when the sample is limited to only high quality studies, there are too few studies reporting information on the pedagogical technique or underlying educational theory to be able to analyze this mechanism further.

Other studies of instructional quality are fewer in number. Studies of **class size and ability tracking** are very limited, and in the case of class size, confounded with additional experiments (the schools whose class sizes were reduced were also recipients of contract teachers). The authors find imprecisely measured effects of 0.05-0.11 standard deviations for math and language respectively (for an 82 to 44 student reduction) and estimate (through instrumental variables) that this effect might be as high as 0.165-0.256 in a scenario in which this confounding did not occur (Duflo, Dupas, and Kremer, 2012). Further, student ability tracking had positive and statistically significant effects on student learning (between 0.156 and 0.166 standard deviations in math and language, respectively), even for students at the lower end of the ability distribution (Duflo, Dupas, and Kremer, 2011). Thus, while mixed ability grouping may still have some peer-to-peer benefits (Garlick, 2013), the benefit to teachers of teaching to smaller and more homogeneous group may have outweighed these peer effects in this particular case.

Further, the pooled effect of **school supplies interventions** is a relatively low 0.022 standard deviations (SE = 0.015, df = 1.25), but the effects of interventions in textbook provision were shown to be particularly high for some sub-groups (high ability students in Kenya (Glewwe, Kremer & Moulin, 2009) and rural students across five West African countries (Frolick and Michaelowa, 2011), though the latter estimates are imprecise. Further, the success of other supply-related interventions were possibly attenuated by implementation issues such as relatively low usage rates of flipcharts in Kenya for the population targeted. Additionally, the effects of one school supply intervention were moderated by household behavioral responses; Das et al (2013) find that unanticipated school supply grants have a slightly greater effect on student performance than anticipated grants, partially due to the fact that households were more likely to reapportion spending away from school supplies in the case of anticipated grants. Finally, while a policy which increased **instructional time** in Ethiopia by 30% is associated with a very large increase in student performance (0.412 standard deviations, SE = 0.184), methodological issues within this paper caution against a strong interpretation.

<u>*Cognitive processing abilities*</u>: As covered above, **school health** interventions have on average some of the lowest and most imprecisely measured pooled effect size estimates in this sample. However, when cognitive outcomes (memory and attention) are assessed, health interventions seem to be more effective (*d* = 0.176, SE = 0.028, df = 2.18). Regarding the heterogeneity of health treatments, I find that malaria prevention/treatment in particular has a sizable pooled effect on cognitive outcomes (*d* = 0.189 standard deviations, SE = 0.022, df = 1.57). Regarding **school meals** (also described above), the average pooled effect size for these

programs is also modest and statistically insignificant, even when examining only cognitive outcomes. The only exception to this is the Vermeersch and Kremer study (2004), which finds insignificant and relatively low effects overall but stronger effects in schools where teachers are more experienced. Finally, even though these school meals programs vary considerably in their length, there is no statistically significant difference in overall pooled effect sizes between programs lasting less than (versus more than) 9 months.

Student or teacher motivation: Programs that provide **student performance incentives** (small monetary incentives to individual students, larger monetary incentives to teams of students, or merit scholarships) to improve student motivation have a high pooled effect size (0.288 standard deviations, SE = 0.015, df = 1), however, this estimate is based on only two studies (containing four treatment arms in total). Meanwhile, interventions that have provided **teacher incentives** have had more mixed results, with an overall pooled effect size of 0.075 (SE = 0.047, p = 0.2). Variation in study effects in this group of studies are attributed to differences in program scale and implementation issues; both Bold et al (2013a) and Bourdon, Frölich, and Michaelowa (2010) suggest that centralized government implementation of teacher contract programs are less effective than more localized/ NGO implementation. Further, Glewwe, Ilias, and Kremer (2010) find that teacher performance incentives produce higher student achievement results when students are tested on the assessments linked to the incentives (but not on other non-linked exams).

School or system accountability: Regarding school or system accountability programs, **information provision interventions** that provide communities with funding information about their schools have large effects on student performance (above 0.33 standard deviations), presumably through the reduction in government capture (Bjorkman, 2006 and Reinikka and

Svennson, 2011), though these impacts may also be due in part to other concurrent government programs (Hubbard 2007). Ngyuen (2008) also finds that information interventions that share information with students about the real returns to education can also lead to improvement in student performance, though providing student school report cards to communities (Piper and Korda, 2011) did not have a strong impact. Finally, **school management interventions** that do not devolve true responsibilities or ownership to school committees tend to lead to relatively low and insignificant impacts, while those that do delegate real responsibilities (hiring/ firing of teachers or putting a monitoring plan in action) have more varied effects, which may potentially be moderated by the level of capacity (measured by literacy rates) of the school management committee itself (Blimpo and Evans, 2011).

Student or community financial restrictions: Programs that help eliminate student or community financial restrictions also have mixed results. **Cost reduction interventions** such as school fee abolition or cash transfers have only moderate overall effects on student performance (pooled effect for all cost reduction interventions = 0.036 standard deviations, SE = 0.036, df = 1.58). However, a school uniform provision program in Kenya had a large performance impact (0.25 standard deviations, significant at 5%) and worked primarily through dramatically increasing attendance rates of students (increase of 44%), though this effect could have been partially due to the psychological effects of being “sponsored” by the implementing NGO. Further, programs that provide schools in under-resourced areas have had extremely varied effects on learning (I^2 = 94.23). Two interventions involving **school infrastructure/ construction** (“girl-friendly schools”) also offered a host of differing complementary interventions, which were not completely implemented in all cases (the program was stalled in Niger). Thus it is difficult to know whether to attribute variation in program effects (high impact in Burkina Faso versus a relatively low impact in Niger) to differences in the implementation of complementary interventions or to differences in initial conditions (the demand for school infrastructure programs appears to have been less in Niger).

Research questions D: Finally, I find that this sample of studies suffers from publication bias, even across multiple sample restrictions (methodologies, topics, and academic disciplines) (for more details, see Chapter V, Section D). However, when I examine the extent of this bias, I calculate that I would need to find 302 studies with no effect at all in order for the average effect size (0.181 standard deviations) to be driven down to an estimate of 0.03 standard deviations, which seems unlikely.

B. Comparison of my findings to other syntheses

The findings of this dissertation are most in line with the findings of McEwan (2013) and Kremer, Brannen, and Glennerster (2013). McEwan (2013) finds that computer-assisted learning and teacher training programs have among the highest pooled effect sizes of studies in his sample. This is consistent with my findings relating to pedagogical methods (both teacher-based and blended learning programs), however I do additionally find high effects for student incentives, though these studies are limited in number. He also finds that intervention types that were among the least effective included school monetary grants and nutritional/ health treatments, which is also in line with my findings, though I do find that health treatments have a relatively large impact on tests of memory and attention. Further, Kremer, Brannen, and Glennerster (2013) find that technology-assisted learning, remedial education, student tracking and the use of contract teachers to be among to most promising interventions. These findings are somewhat in line with mine, though the authors do not include evaluations of classroom-based pedagogy shifts in their review, which had among the highest pooled effect size estimate in my analysis. They do find however, that adaptive learning techniques though computer-assisted learning and after-school targeted remedial education are particularly effective, which is also consist with my findings. However, my findings were quite different than those of Krishnaratne, White & Carpenter (2013), who found that school supplies materials had the highest known impact (but only for Math scores, not Language). Further, when I limit my sample to only randomized controlled trials, for an even stricter comparison to McEwan (2013) and Kremer, Brannen, and Glennerster (2013), my relative findings remain unchanged (see Table 25 below).

Table 25. Pooled effect size of each intervention type (RCTs only)

SAMPLE = RCTs	Estimate	SE	df	(P|t|)	*95% CI.L*	*95% CI.U*	I^2
Quality of Instruction							
School Supplies Provision	0.009	0.003	1	‡	-0.03	0.05	0
Class Size & Composition	0.109	0.049	1	‡	-0.51	0.73	0
Instructional Time†	…	…	…	…	…	…	…
Pedagogical Intervention	1.0**	0.356	13.1	0.015	0.23	1.77	96.1
Student Cognitive Ability							
Health Treatment	-0.008	0.041	3.5	‡	-0.13	0.11	45.7
School Meals/Supplements	0.059	0.022	1.16	‡	-0.15	0.27	0
Student/ Teacher Motivation							
Student Incentives	0.288	0.015	1	‡	0.10	0.48	0
Teacher Incentives	0.0924	0.049	3.9	‡	-0.05	0.23	54.8
School/ System Account.							
Management Intervention	0.016	0.028	3.14	‡	-0.71	0.10	18.7
Information Provision	0.110	0.058	1	‡	-0.63	0.85	14.1
Financial Limitations							
Cost Reduction Intervention	0.114	0.083	1	‡	-0.93	1.16	27.6
Infrastructure + Add. Inputs	0.041	0.019	1	‡	-0.20	0.29	0
OVERALL	0.203	0.059	40.2	0.001	0.08	0.32	89.9

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

† There was only one study which focused on the impact of an increase in instructional time.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted.

C. Limitation of this meta-analysis

There are a number of limitations to this meta-analysis. First, the focus of this dissertation has been on learning outcomes, thus programs that may have had a relatively low impact on learning may still have had a measurable impact on enrollment or health outcomes, for example. I have indicated the occurrence of these additional outcomes within the narrative to an extent, but not in a systematic fashion. Importantly, this dissertation does not argue against the continuation of study in any intervention area, due both to the fact that impacts tend to be heterogeneous (and any additional information can help refine our overall findings) and because certain interventions may have important non-learning impacts, such as the case of unconditional cash transfers in Malawi (Baird, McIntosh, and Ozler, 2011), which had a small effect on learning but large effects on life outcomes.

Secondly, while I attempted to code for implementation fidelity, most studies did not report on this issue, making it difficult to tease apart the effects of poorly run or unimplemented studies from other moderator variables. Further, not all important moderator variables are readily available and quantifiable – such as a country's level of decentralization or the literacy levels of particular communities. In addition, because my sample sizes were so small in some cases (particularly when examining estimates within an intervention type), I was not able to control for certain moderator variables technically but instead provided a narrative review of these studies, attempting to tease out explanations for study variability. And even in cases where I was able to run a meta-regression controlling for a host of moderator variables (i.e. my analysis on pedagogical methods and school health), these moderated estimates are still only correlational in nature and not causal estimates of the effect size differential. In addition, the majority of the evidence in this literature review comes from a set of six countries, which while well dispersed across the continent, are still not representative of the populations of other less-researched countries. However these issues of study geographic dispersion (among others), are limitations of any synthesis of this literature.

A final limitation of the research is that it provides average effect size estimates (outcomes) without accounting for the resource inputs dedicated to each program. That is, while student scholarships may have a higher average impact than school health interventions for example, this finding does not take into account the costs associated with each intervention type. If certain programs are extremely inexpensive, then the fact that their estimated impact may be lower than other interventions, may still make them more cost-effective on average. And while interventions in pedagogical methods are on average only seven to eight months in length (compared to health and economics interventions which last one-two years), their relative cost-effectiveness has not been assessed in this dissertation (nor is this information readily available within the individual studies themselves). Kremer, Brannen, and Glennerster (2013) do find, however, in their cost-effectiveness analysis of a sample of 18 RCTs, that interventions in blended learning and remedial education are among the most cost-effective in their sample. They do not, however, evaluate the cost-effectiveness of teacher pedagogical training interventions (there are none in their sample). Otherwise, they find that contract teachers and the provision of earnings information to be the most cost-effective programs, but their initial effect size estimates for these interventions are larger than my pooled effect size measures, thus it is unclear if these interventions would be as cost-effective in my sample.

Another aspect of cost-effectiveness is the duration of the impact of an intervention on current students, future students, or a school community. All of the interventions in this sample have an impact on student learning in some way (which may or may not persist), though only interventions in pedagogical methods have the potential to affect *how* students learn (in addition to *what* they learn), and this “learning how to learn” process may prove to have long term benefits for students. In addition, pedagogical interventions involving teacher training can potentially affect cohort after cohort of students, producing a multiplier effect on student learning in the long-run. This type of long-term benefit does not necessarily apply to teacher contract, class size reduction, or school meal interventions, for example, as the motivation or ability to improve student performance may decrease with the removal of the these programs. However, this type of long-term effect could also apply to school management trainings or any other capacity-building intervention but is most direct in the case of improving teacher pedagogy. In these ways, the potential cost-effectiveness of pedagogical interventions could be quite high in the long-run.

D. Implications for policy and future research

This final section details the implications of this research on policy, suggests how the design and reporting standards of studies in the fields of education, economics and public health could be improved, and also provides suggestions for the direction of further research in this area.

Policy implications: First, in terms of practical applications of this research, it is clear that there is a large degree of heterogeneity in effect size measures in this sample, even within intervention areas. Because interventions in pedagogical methods have consistently significant and high pooled effect size measures which are robust to sample restrictions and moderator variables, I would encourage the implementation of more programs that either train teachers in innovative pedagogical methods or make use of blended learning techniques. In particular, I would encourage the implementation of programs that employ adaptive learning and teacher coaching. These interventions last nearly half as long as all other interventions in the sample, and they have the potential for multiplier effects, as teachers affect numerous cohorts of students.

Importantly, this meta-analysis highlights the fact that many of the interventions under study work best as complementary interventions to improvements in pedagogical methods. For example, interventions in class size reduction, student tracking, or textbook provision presume that teachers know how to best take advantage of a smaller class size or a more homogenous classroom and that they can integrate these textbooks into everyday instruction. Further, interventions that employ teacher-pay-for performance schemes are partially predicated on the assumption that “newly motivated” teachers know *how* to then improve the performance of their students.

In addition, the effects of a wide variety of interventions in this sample may be more prone to outside influence than interventions in pedagogical methods. For example, successful interventions in teacher contracts were shown to depend on the implementing body, as well as the level of education decentralization. Further, school meal programs and anticipated school supply grants may have unanticipated effects on household behavioral responses (re-apportioning of food or education spending within households). Even school-based management interventions may partially depend on the initial capability levels of parents in the community. In contrast, interventions that are able to change instructional practice or student learning styles may be less vulnerable to these types of external factors. Finally, interventions that reduce the cost of schooling or treat malaria may indeed increase attendance rates or improve cognitive outcomes (test of memory and attention), but they may not always attain their intended effect on academic outcomes if instruction itself is unchanged. For these reasons, I would encourage a renewed focus on improving the quality of instruction itself.

Reporting standards: Regarding recommendations for reporting standards, I begin with a discussion of research in the field of education. Researchers in this field generally provide a detailed theory of change, examine assessment validity and reliability, and are more likely to design experiments that test actual mechanisms (i.e. they ask not “what is the impact of a particular reading program” but instead “what is the impact of using procedural learning strategies in the classroom”). However, a number of these studies do not fully detail the treatment implementation itself. Studies in this field are also less likely to report on student attrition and baseline balance of student characteristics and are also less likely to control for study clustering. In addition, these studies tend to be smaller than studies in the fields of economics and health (40% of studies of pedagogical interventions were conducted with randomized units below 30), and are thus less precisely estimated. Finally, a number of studies in this field were completely excluded from this analysis due to the fact that an effect size could not be extracted from the data given; these studies reported only an ANCOVA table in their results section, without group means or any information on the correlation structure of the variables included in the analysis. Thus my advice for education researchers would be to conduct larger randomized trials and improve study reporting standards to include balance and attrition statistics, as well as sufficient information from which to calculate a standardized effect size with clustered standard errors.

While studies in the field of economics and health have on average better reporting standards in terms of baseline balance and attrition, they often report only fully controlled models, and neglect to report information on the pooled standard deviation of an assessment (thus limiting the ability to both standardize the results and compare them with other studies). These studies often do not report a theory of change regarding learning outcomes and also rarely report reliability and validity statistics of their assessments. I thus recommend that these studies include more unadjusted models, information allowing for the standardization of results, and psychometric measures. Including tests of reliability could improve the discriminatory power of researcher-created tests and result in better quality data. Finally, papers from all fields in this sample do not systematically report on the level of implementation fidelity of each intervention. Including this information could shed much light on the reasons behind differential success between programs.

Future research in this area: As stated earlier, this meta-analysis does not recommend the cessation of impact evaluations on any topic, as there is a limited amount of research on programs that improve learning outcomes in Africa in general (particularly in certain areas/ regions). Further, any additional study in any topical area helps us to better understand what variables moderate these effects overall.

Given the disparate levels of research conducted across contexts and topics, I would first encourage researchers to broaden their geographic focus to countries beyond those six in which the majority of research has been conducted (Kenya, Nigeria, South Africa, Uganda, Burkina Faso, and Madagascar). I would also support the increased study of topics that are pressing to the Sub-Saharan African context such as multi-shift teaching, multi-grade teaching, high stakes testing, language of instruction, and class size, among others (see Chapter V, Section A, Part 1).

The two intervention areas with the largest pooled effect sizes and the most consistently large effects were pedagogical methods and student incentives programs (though the pooled effect of student incentives is based only on two studies with a total of four treatment arms). Thus I would especially support increased experimentation in both of these areas. In particular, within pedagogical methods, the evidence available on the use of different types of instructional techniques or mechanisms (inquiry-based learning versus procedural learning versus conceptual learning etc.) is quite thin; I see this as a priority area. To quote Cohen, Raudenbush, and Ball (2003), I "argue for a model in which the key causal agents are situated in instruction" (pg. 119).

However, this is not the current focus of impact evaluation funding by most international organizations, which have largely funded evaluations in the areas of teacher pay-for-performance, management, and other structural interventions. Moreover, this is has also not been the focus of most international organizations that conduct impact evaluations on the ground in Sub-Saharan Africa. Given the low learning levels of many primary and secondary school students in Sub-Saharan Africa, the potential for very large impacts associated with pedagogical interventions (in both the short and long-term), the fact that many of the other interventions in this sample hinge on teacher performance itself, and the fact that these interventions last on average half as long as other interventions in this sample, I argue that increased investment in programs and evaluations that include adaptive learning, teacher coaching, and a clearly defined instructional mechanism to be tested, would be extremely worthwhile.

REFERENCES

Abeberese, Ama Baafra, Todd J. Kumler, and Leigh L. Linden. (2011). Improving Reading Skills by Encouraging Children to Read: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines. NBER Working Paper No. 17185.

Adedayo, O.A. (1999). Differential Impacts by Gender of Instruction on Achievement in Mathematics at the Tertiary Level. *Educational Studies in Mathematics*, *37*, 83-91.

Adegoke, Benson Adesina. (2011). Effect of Multimedia Instruction on Senior Secondary School Students' Achievement in Physics. *European Journal of Educational Studies*, *3*(3).

Adeleke, M. A. (2007). Strategic improvement of mathematical problem solving performance of secondary school students using procedural and conceptual learning strategies. *Educational Research and Reviews*, *2*(9), 259-263, September.

Agbatogun, Alaba Olaoluwakotansibe. (2012). Exploring the Efficacy of Student Response System in a Sub-Saharan African Country: A Sociocultural Perspective. *Journal of Information Technology Education: Research*, *11*, 249-267.

Ahn, Soyeon, Allison J. Ames, & Nicholas D. Myers. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82*(4), 436-476.

Akinsola, M.K. & I.A. Animasahun. (2007). The Effect of Simulation-Games Environment on Students' Achievement in and Attitudes to Mathematics in Secondary School. *The Turkish Online Journal of Educational Technology – TOJET,* July. ISSN: 1303-6521.

Angrist, Joshua D. and Jorn-Steffen Pischke. (2009). *Mostly Harmless Econometrics*. Princeton, NJ. Princeton University Press.

Angrist, Joshua D. and Jorn-Steffen Pischke. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Economics." *Journal of Economics Perspectives, 24*(2), 3-30.

Awolola, Samuel Adejare. (2011). Effect of brain-based learning strategy on students' achievement in senior secondary school mathematics in Oyo State, Nigeria. *Cypriot Journal of Educational Sciences, 2*, 91-106.

Awoniyi, Adedeji and Florence B. O. Ala. (1985). Effects of Alternative Language Media on Learning in Nigeria. *The Journal of Negro Education*, *54*(2), 225-231.

Baird, Sarah, Craig McIntosh, and Berk Ozler. (2011). Cash or Condition? Evidence from a Cash Transfer Experiment. *The Quarterly Journal of Economics*, *126*, 1709–1753.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235-1264.

Barrera-Osorio, Felipe, Tazeen Fasih, Harry Patrinos, and Lucrecia Santibanez. (2009). *Decentralized Decision-Making in School: Theory and Evidence on School-Based Management*. Washington DC: The World Bank.

Barrow, Lisa and Cecilia Rouse. (2005). Causality, Causality, Causality: The View of Education Inputs and Outputs from Economics. Working Paper, Federal Reserve Bank of Chicago.

WP 2005-15. Prepared for the Consortium for Policy Research in Education, State of Education Policy Research Meeting, February 14-15.

Baumgartner, Jeannine, Cornelius M Smuts, Linda Malan, Jane Kvalsvig, Martha E van Stuijvenberg, Richard F Hurrell, and Michael B Zimmermann. (2012). Effects of iron and n23 fatty acid supplementation, alone and in combination, on cognition in school children: a randomized, double-blind, placebo-controlled intervention in South Africa. *American Journal of Clinical Nutrition, 96*, 1327-38.

Björkman, Martina. (2006). Does Money Matter for Student Performance? Evidence from a Grant Program in Uganda. Innocenzo Gasparini Institute for Economic Research (IGIER), Università Bocconi. Milan, Italy. Working Paper no. 326.

Blimpo, Moussa P. (2010). Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin. New York University. Unpublished Mimeo.

Blimpo, Moussa P. and David K. Evans. (2011). School-Based Management and Educational Outcomes: Lessons from a Randomized Field Experiment. Unpublished Mimeo.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. (2013a). March. Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education. The Center for Global Development, Working Paper no. 321.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, and Justin Sandefur. (2013b). The High Return to Private Schooling in a Low-Income Country. Africa Growth Initiative, Working Paper 5. Brookings Institute.

Bourdon, Jean, Markus Frölich, and Katharina Michaelowa. (2010). Teacher shortages, teacher contracts and their effect on education in Africa. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 173*(Part 1), 93–116.

Bowles, Samuel and Henry M. Levin. (1968). The Determinants of Scholastic Achievement – An Appraisal of Some Recent Evidence. *The Journal of Human Resources, 3*(1), 3-24.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. And Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd, Chichester, UK.

Brown, Byron A. (2010). Teachers' Accounts of the Usefulness of Multigrade Teaching in Promoting Sustainable Human-Development Related Outcomes in Rural South Africa. *Journal of Southern African Studies, 36*(1).

Brooker, Simon, Hellen Inyega, Benson Estambale, Kiambo Njagi, Elizabeth Juma, Caroline Jones, Catherine Goodman, Matthew Jukes. (2013). Impact of malaria control and enhanced literacy instruction on educational outcomes among Kenyan school children: a multi-sectoral, prospective, randomized evaluation. Draft Grantee Final Report. 3ie.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos. (2011). *Making Schools Work: New Evidence on Accountability Reforms*. Washington DC: The World Bank.

Carnoy, Martin and Fabian Arends. (2012). Explaining mathematics achievement gains in Botswana and South Africa. December. *Prospects, 42*(4), 453-468.

Case, Anne and Angus Deaton. (1999). School Inputs and Educational Outcomes in South Africa. *The Quarterly Journal of Economics, 114*(3), 1047-1084.

Cooper, Harris M., Larry V. Hedges & Jeffrey C. Valentine. (Eds.) (2009). *The Handbook of Research Synthesis and Meta-Analysis (2nd Edition).* New York: The Russell Sage Foundation.

Cohen, David, Stephen Raudenbush, and Deborah Loewenberg Ball. (2003). Resources, Instruction, and Research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142.

Cook, Thomas. (2002). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Cook, Thomas & Laura C. Leviton. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, *48*, 449-472.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. (2013). School Inputs, Household Substitution, and Test Scores. *American Economic Journal: Applied Economics, 5*(2), 29–57.

Deaton, Angus. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, *48*(2), 424-55.

Dumitrescu, Anca, Dan Levy, Cara Orfield & Matt Sloan. (2011). Impact Evaluation of Niger's IMAGINE Program. Mathemetica Policy Research. Final Report.

Egger, Mathias, George Davey Smith, Martin Schneider, and Chistoph Minder. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629-634.

Evans, David, Michael Kremer, and Muthoni Ngatia. (2009). "The Impact of Distributing School Uniforms on Children's Education in Kenya," J-PAL mimeo.

Filmer, Deon and Norbert Schady. (2009). School Enrollment, Selection and Test Scores. The World Bank Development Research Group. Human Development and Public Services Team. Impact Evaluation Series No. 34. Policy Research Working Paper 4998.

Fiszbein, Ariel and Norbert Schady. (2009*). Conditional Cash Transfers: Reducing Present and Future Poverty*. A World Bank Policy Research Report.

Fisher, Zachary. (2014). Package 'robumeta' for Robust Variance Estimation. Available at: http://cran.r-project.org/web/packages/robumeta/robumeta.pdf.

Frölich, Markus and Katharina Michaelowa. (2011). Peer effects and textbooks in African primary education*. Labour Economics*, *18*, 474–486.

Garlick, Robert. (2013). Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment. Job Market Paper, University of Michigan.

Gee, Kevin A. (2010). The Impact of School-Based Anti-Malarial Treatment on Adolescents' Cognition: Evidence from a Cluster-Randomized Intervention in Kenya. Doctoral Dissertation. Harvard University, Graduate School of Education.

Glass, Gene V., Barry McGaw, & Mary Lee Smith. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Glass, Gene V. and Mary Lee Smith. (1979). Meta-analysis of research on the relationship of class size and achievement. *Educational Evaluation and Policy Analysis*, *1*(1), 2-16.

Glewwe, Paul, Albert Park, and Meng Zhao. (2011). A Better Vision for Development: Eyeglasses and Academic Performance in Rural Primary Schools in China. J-PAL Working Paper.

Glewwe, Paul, Eric A. Hanushek, Sarah Humpage, and Renato Ravina. (2011). School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010. Working Paper/ Mimeo.

Glewwe, Paul and Eugenie Maïga. (2011). The Impacts of School Management Reforms in Madagascar: Do the Impacts Vary by Teacher Type? J-PAL Working Paper/ Mimeo.

Glewwe, Paul and Michael Kremer. (2006). Schools, Teachers, and Education Outcomes in Developing Countries. *Handbook of Economics of Education, 2*(2), 945-1017.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, *1*(1), 112–135.

Glewwe, Paul, Michael Kremer, Sylvie Moulin and Eric Zitzewitz. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics*, *74*, 251– 268.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. (2010). Teacher Incentives. *American Economic Journal: Applied Economics*, *2*, 205–227.

Gove, Amber and Peter Cvelich. (2011). Early Reading: Igniting Education for All. A report by the Early Grade Learning Community of Practice. Revised Edition. Research Triangle Park, NC: Research Triangle Institute.

Grigorenko, Elena, Robert Sternberg, Matthew Jukes, Katie J. Alcock, Jane Lambo, Damaris Ngorosho, Catherine Nokes, and Donald A. Bundy. (2006). Effects of antiparasitic treatment on dynamically and statically tested cognitive skills over time. *Journal of Applied Developmental Psychology*, *27,* 499–526.

Halpin, Peter F. and Catalina Torrente. (2014). Measuring critical education processes and outcomes: Illustration from a cluster randomized trial in the Democratic Republic of Congo. SREE Spring 2014 Conference Abstract.

Hanushek, Eric A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management*, *1*, 19-41.

Hanushek, Eric A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*, 1141-1177.

Hanushek, Eric A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45-65.

Hanushek, Eric A. (1991). When school finance "reform" may not be a good policy. *Harvard Journal on Legislation*, *28,* 423-456.

Hanushek, Eric A. (1995). Interpreting Recent Research on Schooling in Developing Countries. *The World Bank Research Observer*, *10*(2), 227-246.

Hanushek, Eric A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*, 141-164.

Hanushek, Eric A. and Ludger Woessmann. (2007). The Role of Education Quality for Economic Growth. World Bank Policy Research Working Paper No. 4122.

Hedges, Larry V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370.

Hedges, Larry.V., Elizabeth Tipton, & Matthew C. Johnson. (2010). Jan/Mar. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65.

Hedges, Larry V., Richard D. Laine, Rob Greenwald. (1994). Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student. *Educational Researcher*, *23*(3), 5-14.

Heyneman, Stephen P. & William A. Loxley. (1983). The Effect of Primary School Quality on Academic Achievement across Twenty-nine High and Low-Income Countries. *American Journal of Sociology*, *88*(6), 1162-1194.

Higgins, Julian P T et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343,* d5928.

Higgins, Julian PT and Sally Green (editors). (2011). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Hill, Carolyn J. et al. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, *2*(3), 172–177.

Hubbard, Paul. (2007). Putting the Power of Transparency in Context: Information's Role in Reducing Corruption in Uganda's Education Sector. Center for Global Development, Working Paper Number 136.

Jackson, Gregg.B. (1980). Methods for integrative reviews. *Review of Educational Research*, *50,* 438-460.

Jacoby, Hanan. G. (2002). Is there an intrahousehold ‘flypaper effect’? Evidence from a school feeding programme. *Economic Journal, 112*, 196-221.

Kazianga, Harounan, Damien de Walque, and Harold Alderman. (2012). Educational and Child Labor Impacts of Two Food for Education Schemes: Evidence from a Randomized Trial in Rural Burkina Faso. Working Paper.

Kazianga, Harounan, Dan Levy, Leigh L. Linden, and Matt Sloan. (2013). The Effects of “Girl-Friendly” Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. *American Economic Journal: Applied Economics*, *5*(3), 41–62.

Kellaghan, Thomas and Vincent Greaney. (1992). *Using Examinations to Improve Education: A Study in Fourteen African Countries*. World Bank Technical Paper Number 165; African Technical Department Series.

Kremer, Michael. (2003). Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons. *The American Economic Review, 93*(2), 102-106.

Kremer, Michael and Alaka Holla. (2009). Improving Education in the Developing World: What Have We Learned from Randomized Evaluations? *Annual Review of Economics*, *1*, 513-542.

Kremer, Michael, Conner Brannen & Rachel Glennerster. (2013). The Challenge of Education and Learning in the Developing World. *Science, 340*, 297-299.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. (2009). Incentives to Learn. *The Review of Economics and Statistics*, *91*(3) 437-456.

Krishnaratne, Shari, Howard White, and Ella Carpenter. (2013). Quality education for all children? What works in education in developing countries. 3ie Working Paper, September.

Kurumeh, Seraphina M. S. and Emmanuel. E. Achor. (2008). Effect of Cuisenaire Rods' approach on some Nigeria primary pupils' achievement in decimal fractions. *Educational Research and Review*, *3*(11), 339-343.

Lassibille, Gerard, Jee-Peng Tan, Cornelia Jesse, and Trang Van Nguyen. (2010). Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions. *The World Bank Economic Review*, *24*(2), 303–329.

Light, Richard J. & David B. Pillemer. (1982). Numbers and narrative: Combining their strengths in research reviews. *Harvard Educational Review*, *52*, 1-26.

Light, Richard J., & Paul V. Smith. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Education Review*, *41*, 429-471.

Lipsey, Mark W. & David B. Wilson. (2001). *Practical Meta-Analysis*. Thousand Oaks, Calif: Sage Publications.

Loeb, Susanna & Patrick J McEwan. (2009). Education Policies. Prepared for the NBER Conference on "Targeting Investments in Children: Fighting Poverty When Resources Are Limited." Mimeo.

Lucas, Adrienne M. and Isaac M. Mbiti. (2012). Access, Sorting, and Achievement: The Short-Run Effects of Free Primary Education in Kenya. *American Economic Journal: Applied Economics*, *4*(4), 226–225.

Lucas, Adrienne M., Patrick J. McEwan, Moses Ngware & Moses Oketch. (2013). Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda. Working Paper (September).

Ludwig, Jens, Jeffery R. Kling, and Sendhil Mullainathan. (2011). Mechanism Experiments and Policy Evaluation. *Journal of Economic Perspectives*, *25*(3), 17-38.

Majgaard, Kirsten and Alain Mingat. (2012). *Education in Sub-Saharan Africa: A Comparative Analysis*. The World Bank: Washington DC.

Martinez, Sebastian, Sophie Naudeau, and Vitor Pereira. (2012). The Promise of Preschool in Africa: A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique. The International Initiative for Impact Evaluation (3ie).

McEwan, Patrick. J. (2013). August. Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. Unpublished Manuscript.

Means, Barbara, Yukie Toyama, Robert Murphy, Marianne Bakia & Karla Jones. (2010). Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Washington, D.C.: U.S. Department of Education. Office of Planning, Evaluation, and Policy Development Policy and Program Studies Service.

Nguyen, Trang. (2008). Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar. Working Paper/ Job Market Paper: MIT.

Okoye, Nnamdi and Rose Okechukwa. (2010). The Effect of Concept Mapping and Problem Solving Techniques Strategies on Achievement in Biology among Nigerian Secondary School Students. *Education*, *131*(2).

Olowa, O.W. (2009). Effects of the Problem Solving and Subject Matter Approaches on the Problem Solving Ability of Secondary School Agricultural Education. *Journal of Industrial Teacher Education*, *46*(1).

Onabanjo, Oluyemi I. and P.N. Okpala. (2006). Peer Tutoring – Assisted Instruction, Parent Supportiveness and Student Locus of Control – as Determinants of Academic Achievement in Senior Secondary School Mathematics. Working Paper.

Onabanjo, I. Oluyemi and P.N. Okpala. (2006). Peer Tutoring – Assisted Instruction, Parent Supportiveness and Student Locus of Control – as Determinants of Academic Achievement in Senior Secondary School Mathematics. Unpublished Manuscript, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria.

Onu V. C., M. Eskay, J.N. Igbo, N. Obiyo, and O. Agbo. (2012). Effect of Training in Math Metacognitive Strategy on Fractional Achievement of Nigerian Schoolchildren. *US-China Education Review, B3*, 316-325.

Orkin, Kate. (2013). Lengthening the school day and children’s achievement in Ethiopia. Draft Paper, University of Oxford, April 1.

Pawson, Ray and Nick Tilley. (1997). *Realistic Evaluation*. London and Thousand Oaks, Calif.: Sage Publications.

Petrosino A, Morgan C, Fronius TA, Tanner-Smith EE, Boruch RF. (2012). Interventions in Developing Nations for Improving Primary and Secondary School Enrollment of Children: A Systematic Review. Campbell Systematic Reviews 2012:19.

Piper, Benjamin. (2009). Integrated Education Program: Impact Study of SMRS Using Early Grade Reading Assessment in Three Provinces in South Africa. This publication was produced for review by the United States Agency for International Development. Prepared by RTI International.

Piper, Benjamin and Medina Korda. (2011). EGRA Plus: Liberia, Program Evaluation Report. Produced for review by the United States Agency for International Development, Prepared by RTI International and the Liberian Education Trust.

Plomp, Tjeerd, Jacqueline Pilon, and Ingeborg Janssen Reinen. (1991). The Effectiveness of Remedial Computer Use for Mathematics in a University Setting (Botswana). *Computers and Education*, *16*(4), 337-347.

Postlethwaite, T. Neville. (2004). Monitoring Educational Achievement Fundamentals of educational planning. No. 81. UNESCO. International Institute for Educational Planning.

Prophet, Robert and Nandkishor Badede. (2009). Language and Student Performance in Junior Secondary Science Examinations: The Case of Second Language Learners in Botswana. *International Journal of Science and Mathematics Education*, *7*, 235-251.

Reinikka, Ritva and Jakob Svensson. (2011). The power of information in public services: Evidence from education in Uganda. *Journal of Public Economics*, *95,* 956–966.

Sailors, Misty, James V. Hoffman, P. David Pearson, S. Natasha Beretvas, and Bertus Matthee. (2010). The Effects of First- and Second-Language Instruction in Rural South African Schools, *Bilingual Research Journal, 33*, 21–41.

Sarfo, Frederick K. and Jan Elen. (2007). Developing technical expertise in secondary technical schools: The effect of 4C/ID learning environments. *Learning Environments Research, 10*, 207–221.

Selod, Harris and Yves Zenou. (2003). Private versus public schools in post-Apartheid South African cities: theory and policy implications. *Journal of Development Economics*, *71*, 351– 394.

Sterne, Jonathan. (2010). Meta-analysis in Stata: history, progress and prospects. Department of Social Medicine, University of Bristol, UK. Available from: http://www.stata.com /meeting/10uk/meta_stata.pdf

Talabi, J.K. (1989). The Comparative Effects of Televised and Programmed Instruction. *Journal of Educational Television*, *15*(1), 17-24.

Tipton, Elizabeth. (*in press*). Small sample adjustments for robust variance estimation with meta regression. *Psychological Methods.*

Tipton, Elizabeth. (2014). Effect Sizes for Economics Studies. Working paper. Teachers College, Columbia University.

Tooley, James, Bao Yong, Pauline Dixon, and John Merrifield. (2011). School Choice and Academic Performance: Some Evidence from Developing Countries. *Journal of School Choice*, *5*, 1–39.

USAID. (2013). Results of the 2013 Early Grade Reading and Early Grade Mathematics Assessments (EGRA & EGMA) in Bauchi State. Nigeria Northern Education Initiative (NEI).

UIS-UNESCO Institute for Statistics. (2014, March 30). UIS Statistics. Net Enrollment Ratios. Retrieved from http://data.uis.unesco.org/.

Vegas, Emiliana (ed.). (2005). *Incentives to Improve Teaching: Lessons from Latin America*. Directions in Development series. Washington D.C: The World Bank.

Vermeersch, Christel and Michael Kremer. (2004). School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. Poverty Action Lab Working Paper (also World Bank Policy Research Working Paper No. 3523).

Waddington, Hugh. (2012). How to do a good systematic review of effects in international development: a tool kit. *Journal of Development Effectiveness*, *4*(3), 359–387.

Whaley, Shannon, Marian Sigman, Charlotte Neumann, Nimrod Bwibo, Donald Guthrie, Robert E. Weiss, Susan Alber, and Suzanne P. Murphy. (2003). The Impact of Dietary Intervention on the Cognitive Development of Kenyan School Children. American Society for Nutritional Sciences. *The Journal of Nutrition*, *133*(11 Suppl 2), 3965S-3971S.

Zopluoglu, Cengiz. (2012). A Cross-National Comparison of Intra-Class Correlation Coefficient in Educational Achievement Outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Yaz, 3*(1), 242-278.

APPENDICES

APPENDIX A.
Studies in this Meta-Analysis

[Separate from references above]

Abdu-Raheem, B.O. (2012). Effects of Problem-Solving Method on Secondary School Students' Achievement and Retention in Social Studies in Ekiti State, Nigeria. *Journal of International Education Research, 8*(1), First Quarter.

Ajaja, O. Patrick & Ochuko Urhievwejire Eravwoke. (2010). Effects of Cooperative Learning Strategy on Junior Secondary School Students Achievement in Integrated Science. *Electronic Journal of Science Education*, *14*(1), 1-18.

Baird, Sarah, Craig McIntosh, and Berk Ozler. (2011). Cash or Condition? Evidence from a Cash Transfer Experiment. *The Quarterly Journal of Economics*, *126*, 1709–1753.

Baird, Sarah, Michael Kremer, Edward Miguel, & Joan Hamory Hicks. (2011).Worms at Work: Long-run Impacts of Child Health Gains. Unpublished manuscript.

Barr, Abigail, Frederick Mugisha, Pieter Serneels, and Andrew Zeitlin. (2012). Information and collective action in the community monitoring of schools: Field and lab experimental evidence from Uganda. Unpublished Mimeo.

Baumgartner, Jeannine, Cornelius M Smuts, Linda Malan, Jane Kvalsvig, Martha E van Stuijvenberg, Richard F Hurrell, and Michael B Zimmermann. (2012). Effects of iron and n23 fatty acid supplementation, alone and in combination, on cognition in school children: a randomized, double-blind, placebo-controlled intervention in South Africa. *American Journal of Clinical Nutrition, 96*, 1327-38.

Bimbola, Oludipe and Oludipe I. Daniel. (2010). Effect of constructivist-based teaching strategy on academic performance of students in integrated science at the junior secondary school level. *Educational Research and Reviews, 5*(7), 347-353.

Björkman, Martina. (2006). Does Money Matter for Student Performance? Evidence from a Grant Program in Uganda. Innocenzo Gasparini Institute for Economic Research (IGIER), Università Bocconi. Milan, Italy. Working Paper no. 326.

Blimpo, Moussa P. (2010). Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin. New York University. Unpublished Mimeo.

Blimpo, Moussa P. and David K. Evans. (2011). School-Based Management and Educational Outcomes: Lessons from a Randomized Field Experiment. Unpublished Mimeo.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. (2013a). March. Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education. The Center for Global Development, Working Paper no. 321.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, and Justin Sandefur. (2013b). The High Return to Private Schooling in a Low-Income Country. Africa Growth Initiative, Working Paper 5. Brookings Institute.

Bourdon, Jean, Markus Frölich, and Katharina Michaelowa. (2010). Teacher shortages, teacher contracts and their effect on education in Africa. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 173*(Part 1), 93–116.

Brooker, Simon, Hellen Inyega, Benson Estambale, Kiambo Njagi, Elizabeth Juma, Caroline Jones, Catherine Goodman, Matthew Jukes. (2013). Impact of malaria control and enhanced literacy instruction on educational outcomes among Kenyan school children: a multi-sectoral, prospective, randomized evaluation. Draft Grantee Final Report. 3ie.

Clarke, Siân E., Matthew C H Jukes, J Kiambo Njagi, Lincoln Khasakhala, Bonnie Cundill, Julius Otido, Christopher Crudder, Benson B A Estambale, and Simon Brooker. (2008). Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: a cluster randomized, double-blind, placebo-controlled trial. *The Lancet*, *372*, 127–138.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman. (2013). School Inputs, Household Substitution, and Test Scores. *American Economic Journal: Applied Economics, 5*(2): 29–57.

DSD, SASSA and UNICEF. (2012). The South African Child Support Grant Impact Assessment: Evidence from a survey of children, adolescents and their households. Pretoria: UNICEF South Africa.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, *101*(August), 1739–1774.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. (2012). School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Public Schools. NBER Working Paper Series. No. 17939.

Dumitrescu, Anca, Dan Levy, Cara Orfield & Matt Sloan. (2011). Impact Evaluation of Niger's IMAGINE Program. Mathemetica Policy Research. Final Report.

Evans, David, Michael Kremer, and Muthoni Ngatia. (2009). “The Impact of Distributing School Uniforms on Children’s Education in Kenya,” J-PAL mimeo.

Frölich, Markus and Katharina Michaelowa. (2011). Peer effects and textbooks in African primary education. *Labour Economics, 18*, 474–486.

Gee, Kevin A. (2010). The Impact of School-Based Anti-Malarial Treatment on Adolescents' Cognition: Evidence from a Cluster-Randomized Intervention in Kenya. Doctoral Dissertation. Harvard University, Graduate School of Education.

Githau, Bernard. N. and Rachel Angela Nyabwa. (2008). Effects of Advance Organiser Strategy during Instruction on Secondary School Students’ Mathematics Achievement in Kenya’s Nakuru District. *International Journal of Science and Mathematics Education*, *6*, 439-457.

Glewwe, Paul and Eugenie Maïga. (2011). The Impacts of School Management Reforms in Madagascar: Do the Impacts Vary by Teacher Type? J-PAL Working Paper/ Mimeo.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, *1*(1), 112–135.

Glewwe, Paul, Michael Kremer, Sylvie Moulin and Eric Zitzewitz. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics, 74,* 251– 268.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. (2010). Teacher Incentives. *American Economic Journal: Applied Economics*, *2*, 205–227.

Grigorenko, Elena, Robert Sternberg, Matthew Jukes, Katie J. Alcock, Jane Lambo, Damaris Ngorosho, Catherine Nokes, and Donald A. Bundy. (2006). Effects of antiparasitic treatment on dynamically and statically tested cognitive skills over time. *Journal of Applied Developmental Psychology*, *27,* 499–526.

Jinabhai, Champak, Myra Taylor, Anna Coutsoudis, Hoosen M. Coovadia, Andrew M. Tompkins, and Keith R. Sullivan. (2001). A randomized controlled trial of the effect of antihelminthic treatment and micronutrient fortification on health status and school performance of rural primary school children. *Annals of Tropical Paediatrics, 21*, 319–333.

Jukes, Matthew, Margaret Pinder, Elena L. Grigorenko, Helen Bañnos Smith, Gijs Walraven, Elisa Meier Bariau, Robert J. Sternberg, Lesley J. Drake, Paul Milligan, Yin Bun Cheung, Brian M. Greenwood, and Donald A. P. Bundy. (2006). Long-Term Impact of Malaria Chemoprophylaxis on Cognitive Abilities and Educational Attainment: Follow-Up of a Controlled Trial. *PLOS Clinical Trials*, August, e19.

Kazianga, Harounan, Damien de Walque, and Harold Alderman. (2012). Educational and Child Labor Impacts of Two Food for Education Schemes: Evidence from a Randomized Trial in Rural Burkina Faso. Working Paper.

Kazianga, Harounan, Dan Levy, Leigh L. Linden, and Matt Sloan. (2013). The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. *American Economic Journal: Applied Economics*, *5*(3), 41–62.

Kiboss, Joel Kipkemboi. (2012). Effect of Special E-Learning Program on Hearing-Impaired Learners' Achievement and Perception of Basic Geometry in Lower Primary *Mathematics. Journal of Educational Computing Research*, *46*(1), 31-59.

Korsah, G. Ayorkor, Jack Mostow, M. Bernardine Dias, Tracy Morrison Sweet, Sarah M. Belousov, M. Frederick Dias, and Haijun Gong. (2010). Improving Child Literacy in Africa: Experiments with an Automated Reading Tutor. *Information Technologies & International Development*, *6*(2), 1–19.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. (2009). Incentives to Learn. *The Review of Economics and Statistics*, *91*(3) 437-456.

Lassibille, Gerard, Jee-Peng Tan, Cornelia Jesse, and Trang Van Nguyen. (2010). Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions. *The World Bank Economic Review, 24*(2), 303–329.

Louw, Johann, Johan Muller, and Colin Tredoux. (2008). Time-on-task, technology and mathematics achievement. *Evaluation and Program Planning, 31*, 41–50.

Lucas, Adrienne M. and Isaac M. Mbiti. (2012). Access, Sorting, and Achievement: The Short-Run Effects of Free Primary Education in Kenya. *American Economic Journal: Applied Economics*, *4*(4), 226–225.

Lucas, Adrienne M., Patrick J. McEwan, Moses Ngware & Moses Oketch. (2013). Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda. Working Paper (September).

Martinez, Sebastian, Sophie Naudeau, and Vitor Pereira. (2012). The Promise of Preschool in Africa: A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique. The International Initiative for Impact Evaluation (3ie).

Miguel, Edward and Michael Kremer. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica, 72*(1), 159-217.

Nguyen, Trang. (2008). Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar. Working Paper/ Job Market Paper: MIT.

Nwagbo, Chinwe. (2006). Effects of two teaching methods on the achievement in and attitude to biology of students of different levels of scientific literacy. *International Journal of Educational Research*, *45*, 216–229.

Orkin, Kate. (2013). Lengthening the school day and children's achievement in Ethiopia. Draft Paper, University of Oxford. April 1.

Piper, Benjamin. (2009). Integrated Education Program: Impact Study of SMRS Using Early Grade Reading Assessment in Three Provinces in South Africa. This publication was produced for review by the United States Agency for International Development. Prepared by RTI International.

Piper, Benjamin and Medina Korda. (2011). EGRA Plus: Liberia, Program Evaluation Report. Produced for review by the United States Agency for International Development, Prepared by RTI International and the Liberian Education Trust.

Reinikka, Ritva and Jakob Svensson. (2011). The power of information in public services: Evidence from education in Uganda. *Journal of Public Economics*, *95*, 956–966.

Sailors, Misty, James V. Hoffman, P. David Pearson, S. Natasha Beretvas, and Bertus Matthee. (2010). The Effects of First- and Second-Language Instruction in Rural South African Schools, *Bilingual Research Journal, 33,* 21–41.

Spratt, Jennifer, Simon King, and Jennae Bulat. (2013). June. Independent Evaluation of the Effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) Program in Mali. RTI International. Endline Report.

Tooley, James, Bao Yong, Pauline Dixon, and John Merrifield. (2011). School Choice and Academic Performance: Some Evidence From Developing Countries. *Journal of School Choice*, *5*, 1–39.

Van Staden, Annalene. (2011). Put reading first: Positive effects of direct instruction and scaffolding for ESL learners struggling with reading. *Perspectives in Education*, *29*(4), 10-21.

van Stuijvenberg, Elizabeth M., Jane D Kvalsvig, Mieke Faber, Marita Kruger, Diane G Kenoyer, and AJ Spinnler Benadé. (1999). Effect of iron-, iodine-, and b-carotene–fortified biscuits on the micronutrient status of primary school children: a randomized controlled trial. *American Journal of Clinical Nutrition*. *69*, 497–503.

Vermeersch, Christel and Michael Kremer. (2004). School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. Poverty Action Lab Working Paper (also World Bank Policy Research Working Paper No. 3523).

Wachanga, Samuel W. & John Gowland Mwangi. (2004). Effects of the Cooperative Class Experiment Teaching Method on Secondary School Students' Chemistry Achievement in Kenya's Nakuru District. *International Education Journal*, *5*(1).

Whaley, Shannon, Marian Sigman, Charlotte Neumann, Nimrod Bwibo, Donald Guthrie, Robert E. Weiss, Susan Alber, and Suzanne P. Murphy. (2003). The Impact of Dietary Intervention on the Cognitive Development of Kenyan School Children. American Society for Nutritional Sciences. *The Journal of Nutrition*, *133*(11 Suppl 2), 3965S-3971S.

APPENDIX B.
Systematic Search Documentation

A. Electronic Bibliographic Search

1- Meta/Aggregated-Databases (Searched from January 1980 – February 2013)
I searched the following sources which pull their search results from multiple databases/ journals:

a. Columbia University "article" search through Columbia libraries; this search engine searches all databases for studies relevant to my search terms (this search is meant to catch any study I may have missed in more narrow/ targeted searches).
b. CU custom search - This feature searches all databases/journals by field (I chose Education, Economics, International/Area Studies, and Political Science/International Affairs).
c. ERIC – through EBSCO host (searches multiple databases at once – searched for both Education and Economics journals)

2- Individual Databases (Searched from January 1980 – February 2013)

a. ERIC
b. Econlit Full text
c. Education Full Text (H.W. Wilson)
d. Education Research Complete
e. JSTOR
f. Academic Search Complete
g. Business Source Complete
h. Social Sciences Full Text (H.W. Wilson)

3- Individual Journals (Searched from January 1980 – February 2013)

	Journal
1	*AEA: American Economic Journal: Applied Economics*
2	*AEA: American Economic Journal: Economic Policy*
3	*AEA: American Economic Journal: Economic Policy (POL)*
4	*AEA: American Economic Journal: Macroeconomics*
5	*AEA: American Economic Journal: Microeconomics*
6	*AEA: American Economic Review*
7	*AEA: Journal of Economic Literature (JEL)*
8	*AEA: Journal of Economic Perspectives (JEP)*
9	*African Journal of Educational Studies in Mathematics and Sciences*
10	*Early Childhood Research Quarterly*
11	*Econometrica*
12	*Economics of Education Review*
13	*Education Evaluation and Policy Analysis*
14	*International Journal of Educational Research*

15	*International Journal of Science and Mathematics Education*
16	*Journal of Development Economics*
17	*Journal of Educational Psychology*
18	*Journal of Human Resources*
19	*Journal of International Education Research*
20	*Journal of Policy Analysis & Management*
21	*Journal of Public Economics*
22	*NBER - WORKING PAPERS*
23	*Quarterly Journal of Economics*
24	*Review of Educational Research*
25	*World Bank Economic Review*

B. Research Center Publications (Searched from January 1980 – April 2013)

1	3ie
2	Brookings Institute
3	Center for Global Development
4	Centre for Economic Policy Reseach, UK - Discussion Papers
5	Innocenzo Gasparini Institute for Economic Research (IGIER)
6	Innovations for Poverty Action (based at Yale University)
7	Institute for Fiscal Studies, London
8	Institute for the Study of Labor (Germany).
9	Institute of Education, University of London
10	Mathematica
11	MIT Poverty Action Lab (J-PAL)
12	National Center for Education Evaluation and Regional Assistance (U.S. Department of Education)
13	National Center on Performance Incentives (Nashville, TN)
14	Population Aging Research Council
15	RAND
16	Research Triangle Institute
	THE WORLD BANK:
17	WB- AIM
18	WB- DEC (DIME)
19	WB- PREM (external and internal)
20	WB- The World Bank Impact Evaluation Working Paper Series
21	WB- World Bank Policy Research Working Paper Series
22	World Bank Open Knowledge Repository

C. Citation Tracking or "Snowballing" Technique
The following books and articles were found to be particularly influential and comprehensive and were thus searched systematically for impact evaluations in Africa (references searched in their entireties)

1- Barrera-Osorio, Felipe et al. (2009). *Decentralized Decision-Making in Schools: The Theory and Evidence on School-Based Management*. The World Bank, Directions in Development.

2- Fiszbein, Ariel and Norbert Schady. (2009*). Conditional Cash Transfers: Reducing Present and Future Poverty*. A World Bank Policy Research Report.

3- Kremer, Michael (2003). Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons. *The American Economic Review*. Vol. 93, No. 2, pg. 102-106.

4- Kremer, Michael and Alaka Holla (2009). Improving Education in the Developing World: What Have We Learned from Randomized Evaluations? *Annual Review of Economics*. 1: 513–42.

5- Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos. (2011). *Making Schools Work: New Evidence on Accountability Reforms*. The World Bank, Washington D.C.

6- Fiszbein, Ariel; Schady, Norbert; Ferreira, Francisco H. G.; Grosh, Margaret; Keleher, Niall; Olinto, Pedro; Skoufias, Emmanuel. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty*. © Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/2597

7- Barrera-Osorio, Felipe, Harry Anthony Patrinos, and Quentin Wodon (eds.). (2009). *Emerging Evidence on Vouchers and Faith-Based Providers in Education: Case Studies from Africa, Latin America, and Asia*. © Washington, DC: World Bank.

8- Glewwe, Paul and Edward Miguel. (2008). The Impact of Child Health and Nutrition on Education in Less Developed Countries (Chapter for the *Handbook of Development Economics*, Vol. 4).

9- Eilander, Ans et al. (2010). Multiple micronutrient supplementation for improving cognitive performance in children: systematic review of randomized controlled trials. *American Journal of Clinical Nutrition*; 91:115–30.

D. Conference or Workshop Presentations
I tracked references from two major World Bank conferences pertaining to my research questions, as well as references from the Economics and Education workshop held weekly at Teachers College, Columbia University (attended from September 2009 – February 2013):

1- "What Works in Education - Policy Research Colloquium"
Location: The World Bank
Date held: 04.29.2011
Website:
http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTEDUCATION/0,,contentMDK:22896219~menuPK:282428~pagePK:64020865~piPK:51164185~theSitePK:282386,00.html

2- "Workshop on Equity, Development & Policy: Evidence, New Ideas & Future Directions."
Location: The World Bank
Date held: 06.10.2011

Website:
http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/0,,contentMDK:22930707~pagePK:210058~piPK:210062~theSitePK:336992,00.html

3- "School Feeding Programs: Evidence and Policy Lessons"
Location: International Food Policy Research Institute.
Date held: 09.01.2009
Website:
http://www.ifpri.org/event/school-feeding-programs-evidence-and-policy-lessons

E. Websites/ CVs of influential authors
The websites and full body of work (both published and unpublished studies) of key researchers in this field were reviewed for additional papers:
1- Michael Kremer (Harvard University)
2- Esther Duflo (MIT)
3- Pascaline Dupas (Stanford University)
4- Ted Miguel (University of California at Berkeley)

F. Contacting individual researchers
1- An e-mail was sent to World Bank staff (June 2012) in the education research group asking for suggestions of any missing citations (from the list of education impact evaluations available at the time in Africa).
2- Suggestions of additional citations to include in the research plan were received from World Bank staff (education research group): Harry Patrinos (Manager, Education, Human Development Network), Oni Lusk-Stover (Operations Officer, Education, the Human Development Network) & Emilio Porta (Senior Education Specialist, Education, Human Development Network).

Search Terms:

Methodological terms: _(AB "impact evaluation" OR AB impact OR AB evaluat* OR AB "program evaluation" OR AB "experiment*" OR AB "quasi-experiment*" OR AB "quasi experiment*" OR AB "exogenous" OR AB random* OR AB "evidence" OR AB "effect*" OR TI "affect*" OR "Quasiexperimental Design" OR "Program Evaluation" OR "Program Effectiveness" OR "Data Analysis" OR "Evaluation Methods" OR "Evaluation Research" OR "School Effectiveness" OR "Evaluation" OR "Experiment*" OR "Intervention" OR "Economic Research" OR "Outcomes of Education" OR "Student Evaluation" OR "Experimental Group*" OR "Treatment Group*" OR "Control Group*" OR "pretest*" OR "posttest*" OR "Pretests Posttests") AND_
Education terms: _(AB educat* OR AB school* OR AB learn* OR AB teach* OR "International Education" OR "Educational Development" OR "Comparative Education") AND_
Sub-Saharan Africa terms: _(AB "Sub-Saharan Africa*" OR AB "Algeria*" OR AB "Angola*" OR AB "Benin*" OR AB "Botswana*" OR AB "Burkina Faso*" OR AB "Burundi*" OR AB "Cameroon*" OR AB "Canary Islands" OR AB "Cape Verde" OR AB "Central African Republic*" OR AB "Chad*" OR AB "Comoros" OR AB "Cote d'Ivore" OR AB "Democratic Republic of Congo" OR AB "Djibouti*" OR AB "Egypt*" OR AB_

"Equatorial Guinea*" OR AB "Eritrea*" OR AB "Ethiopia*" OR AB "Gabon*" OR AB "Gambia*" OR AB "Ghana*" OR AB "Guinea*" OR AB "Guinea-Bissau*" OR AB "Kenya*" OR AB "Lesotho*" OR AB "Liberia*" OR AB "Libya*" OR AB" Madagascar" OR AB "Malawi*" OR AB "Mali*" OR AB "Mauritania*" OR AB "Mauritius" OR AB "Morocco*" OR AB "Mozambique" OR AB "Namibia*" OR AB "Niger*" OR AB "Nigeria*" OR AB "Republic of the Congo" OR AB "Rwanda*" OR AB "Sao Tome and Principe" OR AB "Senegal*" OR AB "Seychelles" OR AB "Sierra Leone*" OR AB "Somalia*" OR AB "South Africa*" OR AB "Sudan*" OR AB "Swaziland*" OR AB "Tanzania*" OR AB "Togo*" OR AB "Tunisia*" OR AB "Uganda*" OR AB "Western Sahara*" OR AB "Zambia*" OR AB "Zimbabwe")AND
Intervention-specific search terms:
UPE OR "universal primary education" OR "abolish* school fees" OR "eliminat* school fees" OR "school fees" OR EFA
"class size" OR "pupil teacher ratio" OR "Teacher Student Ratio" OR "student teacher ratio" OR "teacher student ratio"
"equity" OR "holistic" or "multi-sector*"
CCT OR "conditional cash transfer*" OR "cash transfer*"
AB "information" OR AB "accountab*"
private OR "private education" OR "private provision" OR "private provider*" OR "private school*"
health OR nutrition OR worm* OR vitamins* OR iodine OR meal* OR "school feeding"
"school resources" OR "textbook*" OR "Educational Resources" OR "school grants"
"school based management" OR decentraliz* OR "community school*" OR "community-managed school*" OR "Institutional Characteristics" OR "Parent Associations" OR "Parent Teacher Cooperation" OR "Parents" OR "Community Involvement" OR "Citizen Participation" OR "School Community Relationship" OR "Partnerships in Education" OR "Parent Role"
scholarship* OR incentive* OR "Merit Scholarships" OR "teacher incentive*" OR "teacher contract*" OR bonus OR "pay for performance" OR "performance pay" OR "Beginning Teachers" OR "Contract Training" OR "Student Teachers"
"teacher training" OR "certification" OR "qualification*" OR "Teacher Education" OR "Teacher Improvement" OR "Preservice Teachers" OR "Inservice Teacher Education" OR "Teaching Experience" OR "Teacher Competencies" OR "Teacher Characteristics" OR "Teacher Qualifications" OR "Faculty Development" OR "teacher professional development" OR "professional development" OR "Pedagogical Content Knowledge" OR "Teacher Educator Education" OR "Teacher Education Programs" OR "Teacher Effectiveness"
tracking OR "peer effects" OR "Peer Influence" OR "Track System (Education)" OR "Heterogeneous Grouping" OR "Homogeneous Grouping" OR "Ability Grouping"
"high stakes testing" OR "High Stakes Tests" OR "accountability system" or "national exams"
"distance education" OR "IT" OR "ICT" OR "Information Technology" OR "technolog*" OR "interactive instruction" OR "Electronic Learning" OR "Educational Games" OR "Computer Assisted Instruction" OR "Educational Technology" OR "Educational Radio" OR "Online Courses" OR "Web Based Instruction" OR "Blended Learning" OR "Technology Integration" OR "Computer Uses in Education" OR "Access to Computers" OR "Handheld Devices" OR "Technology Uses in Education"

"instructional time" OR "length of school day" or "days in school"
"school choice" OR "voucher*"
"infrastructure" OR "school renovation" OR "school repairs" OR "school building" OR "School Construction" OR "Construction Programs" OR "Educational Facilities"
"tutor*" OR "supplemental instruction" OR "SI" OR "after school program" OR "additional instruction" OR "summer program" OR "peer tutor*" OR "private tutor*"
"instructional techniques" OR "pedagogy" OR "pedagogical technique*" OR "teaching technique*" OR "method of instruction" OR "Curriculum Development" OR "Instructional Materials" OR "Teaching Methods" OR "Educational Strategies" OR "Learning Strategies" OR "Instructional Effectiveness" OR "Instructional Design" OR "Critical Thinking" OR "Educational Innovation" OR "Instructional Development" OR "Educational Practices" OR "Thinking Skills" OR "Lecture Method" OR "Creative Teaching" OR "Teaching Styles" OR "Classroom Techniques" OR "Progressive Education"
"Language of Instruction" OR "Native Language" OR "English (Second Language)" OR "Second Language Learning" OR "Language Planning" OR "Language Policy" OR "African Language*" OR "Language Usage" OR "Second Language Instruction" OR "Language Processing" OR "Language Acquisition" OR "biligual teaching" or bilingual
"boarding" OR "day school" OR "female-only" OR "girls school" OR "all-girls" OR "all-female" OR "single sex" OR "co-educational" OR "co-ed" OR "coed" OR "school type" OR "type of school" OR "religious school*" OR "Womens Education"
"multi-grade" OR "shift teaching" OR "multigrade" OR double-shift OR "double shift"
"Beginning Reading" OR "Reading Tests" OR "Reading Skills" OR "Literacy" OR "Reading Diagnosis" OR "Reading Instruction" OR "Reading Achievement" OR "Mathematics Achievement" OR "Remedial Instruction" OR "Numeracy"

APPENDIX C.
Mathematical Appendix

In all studies, the goal is to estimate $\delta_t = (\mu_t - \mu_c)/\sigma_t$ using some estimate d_t, as well as its sampling variance, $V(d_t)$. Importantly, in some studies, δ_w is instead estimated. When this is the case, I refer to the following conversion (Hedges, 2007),

$$d_t = d_w * \sqrt{(1 - \rho)}$$

$$V\{d_t\} = (1 - \rho) * V\{d_w\}$$

where d_w and $V\{d_w\}$ are the estimate of δ_w and its sampling variance respectively. Here ρ is the ICC.

I. Cases with standardized data

If a study first standardized Y, it did so by subtracting the mean and dividing by the standard deviation of the total, s_t. Based on results from Hedges (2007), in the regression table, this means that we see,

$$b_{s1} = \frac{\bar{y}_t - \bar{y}_c}{s_t}$$

and a consistent estimator of δ_t is therefore

$$d_{t2} = b_{s1}\sqrt{1 - \frac{2(n-1)\rho}{N_T + N_C - 2}},$$

with variance

$$V\{d_{t2}\} = \left(\frac{N^T + N^C}{N^T N^C}\right)(1 + (n-1)\rho) + \delta_T^2 \left(\frac{(N-2)(1-\rho)^2 + n(N-2n)\rho + 2(N-2n)\rho(1-\rho)}{2(N-2)[(N-2) - 2(n-1)\rho]}\right)$$

where d_{t2} is the second effect size estimator of δ_t given in Hedges (2007).

II. Conversions for primary studies not accounting for clustering

In some studies, students were clustered in schools, but this clustering was not taken into account in the estimation of the treatment impact in the primary study. This means that the reported standard error in the primary paper is incorrect and often too small. In order to adjust these and properly take into account this clustering, I use the following result from Hedges (2007),

$$V\{d_w\} = \left(\frac{N^T + N^C}{N^T N^C}\right)\left(\frac{1 + (\tilde{n} - 1)\rho}{1 - \rho}\right) + \left(\frac{\delta_W^2}{2(N - M)}\right)$$

$$where\ \tilde{n} = \frac{N^C \sum_{i=1}^{m^T}(n_i^T)^2}{N^T N} + \frac{N^T \sum_{i=1}^{m^C}(n_i^C)^2}{N^C N}$$

Here N^T is equal to the total number of units in the treatment groups, N^C is equal to the total number of units in the control group, ñ is equal to the number of students in the classroom or school (unit that is subject to clustering), ρ is a measure of the inter-class correlation, which is both country and subject-specific (where available) and found in both the TIMMS and SACMEQ datasets (see Postlethwaite 2004 & Zopluoglu 2012), δ_w^2 is the estimate (squared) of the effect size (estimated using the amount of variation within studies), *N* is the total number of student in the sample, and *M* is the total number of treatment and control groups.

Note that ñ is a weighted average calculated with n^T (average number of students in the treatment classes/ schools), n^C (average number of students in the control classes/ schools), m^T (number of treatment group), and m^C (number of control groups).

III. Conversions for studies only reporting cluster robust standard errors

In some studies, the outcomes were not standardized before analysis and no measure of variability (σ^2) was reported in the paper. Based on results from Tipton (2014), however, when robust or cluster robust standard errors were reported, I was able to extract the necessary information for estimating δ_t as follows.

Assume a cluster randomized study with:

- $n_j = n$ units in level 1 (e.g., students)
- m_t treatment schools, m_c control schools
- $N_t = m_t * n$ total treatment students, $N_c = m_c * n$ total control students
- $\rho = \sigma_B^2/\sigma_T^2$ intra-class correlation

Then in the regression,

$$b_1 = \bar{Y}_T - \bar{Y}_C$$

which is an unstandardized measure of the treatment impact. The cluster robust estimator of $V(b_1)$ is

$$v(b_1) = l'(\mathbf{X'X})^{-1}\Sigma \mathbf{X}_j' \mathbf{e}_j \mathbf{e}_j' \mathbf{X}_j (\mathbf{X'X})^{-1} l$$

where $l = (0,1)$ and $\mathbf{e}_j = \mathbf{Y}_j - \mathbf{X}_j\mathbf{b}$ are the observed residuals in cluster *j*. Note that in control schools, $\mathbf{e}_j = \mathbf{Y}_j - \bar{Y}_c \mathbf{1}$, and in treatment schools, $\mathbf{e}_j = \mathbf{Y}_j - \bar{Y}_t \mathbf{1}$.

Thus the following estimate

$$d_T = \frac{b_1}{\sqrt{v(b_1)}} \sqrt{\left(\frac{1+(n-1)\rho}{n}\right) + \left(\frac{m_t-1}{m_t^2} + \frac{m_c-1}{m_c^2}\right)}$$

is a consistent estimate of $\delta_T = \beta_1/\sigma_T$ and that

$$\mathrm{V}(d_\mathrm{T}) = \left(1+(n-1)\rho\right)\left[\left(\frac{N_t+N_c}{N_tN_c}\right) + \frac{\delta_T^2}{2n}\left(\frac{\frac{m_t-1}{m_t^4}+\frac{m_c-1}{m_c^4}}{\frac{m_t-1}{m_t^2}+\frac{m_c-1}{m_c^2}}\right)\right]$$

IV. Conversions for studies only reporting robust standard errors (non-clustered)

Assume the same set-up as above. Here the non-clustered robust standard error estimator of $\mathrm{V}(b_1)$ is

$$v(b_1) = l'(\mathbf{X'X})^{-1}\Sigma\mathbf{X}_j'\mathbf{E}_j'\mathbf{X}_j(\mathbf{X'X})^{-1}l$$

where $l = (0,1)$ and $\mathbf{E}_j = \mathrm{diag}(e_{1j}^2, e_{2j}^2, \ldots, e_{nj}^2)$, and $e_{ijt}^2 = (Y_{ij} - \overline{Y}_t)^2$ and $e_{ijc}^2 = (Y_{ij} - \overline{Y}_c)^2$ are the observed residuals in cluster *j*.

Then the estimator

$$d_T = \frac{b_1}{\sqrt{v(b_1)}} \sqrt{\left(\frac{N_t-1}{N_t^2} + \frac{N_c-1}{N_c^2}\right) - (n-1)\rho\left(\frac{1}{N_t^2} + \frac{1}{N_c^2}\right)}$$

is a consistent estimate of $\delta_T = \beta_1/\sigma_T$ and

$$V(d_t) = (1+(n-1)\rho)\left(\frac{N_t+N_c}{N_tN_c}\right) + \frac{\delta_T^2}{2}\left(\frac{c}{e}\right)$$

is an estimator of the variance of d_T.

V. Conversions for instrumental variables results

One study reported results from an IV analysis, where the IV was a continuous variable. Here the measure of the impact was an estimated regression coefficient, b_1 and its variance $v(b_1)$, found in a regression table.

In this case, Tipton (2014) shows that we can convert this effect (b_1) to a standardized mean difference in two steps. First, note that a regression coefficient can be converted to a correlation using,

$$r = b_1\left(\sqrt{\frac{\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2}}\right) = b_1 \frac{SD(X)}{SD(Y)}$$

where SD(X) and SD(Y) are the standard deviations of X and Y respectively. Second, we can convert r to d using Cohen (1986),

$$d = \frac{2r}{\sqrt{(1-r^2)}}$$

where *r* is related to b_1 from above. Finally, Tipton provides the following delta-method estimator of the variance of *d*,

$$V(d) \approx = \left[\frac{4}{(1-r^2)^3}\right]\left[\frac{V(X)}{V(Y)}\right]V(b_1)$$

APPENDIX D.
Quality Index

QUALITY INDEX MEASURE	**[Score]**			
Randomized Controlled Trials	**0 points**	**1 point**	**2 points**	Total*
Intervention/ Experiment Description	Description vague or unclear	Adequate description	Detailed description	2
Presentation	Very poor presentation/ poor language/ tables mis-labled	Adequate (poor language possible, but data tables are clear)	Professional quality presentation	2
Balance	Balance of groups not checked	Balance of groups checked (limited check on few variables)	Balance of groups fully checked (and any imbalance accounted for technically)	2
Attrition	Attrition within program is not stated	Attrition within program is stated	Attrition within program is stated and addressed	2
Spillovers (did not incorporate in index b/c not applicable to every study)	*No information regarding any spillovers reported*	*Spillovers reported*	*Spillovers reported & addressed*	*0*
OVERALL				**8**
Instrumental Variables	**0 points**	**1 point**	**2 points**	Total
Intervention/ Experiment Description	Description vague or unclear	Adequate description	Detailed description	2
Presentation	Very poor presentation/ poor language/ tables mis-labled	Adequate (poor language possible, but data tables are clear)	Professional quality presentation	2
Quality of 1SLS & 2SLS	reasoning provided re: exclusion restriction	test for first stage (positive correlation) + reasoning re: exclusion restriction	test for first stage + reasoning re: exclusion restriction + tests of exclusion restriction	2
OVERALL				**6**
Difference in Difference	**0 points**	**1 point**	**2 points**	Total
Intervention/ Experiment Description	Description vague or unclear	Adequate description	Detailed description	2
Presentation	Very poor presentation/ poor language/ tables mis-labled	Adequate (poor language possible, but data tables are clear)	Professional quality presentation	2
Balance/ Trends	Balance/ trends of groups not checked	Balance/ trends of groups checked (limited check on few variables)	Balance/ trends of groups fully checked (and any imbalance accounted for technically)	2
OVERALL				**6**

[Appendix D continued]

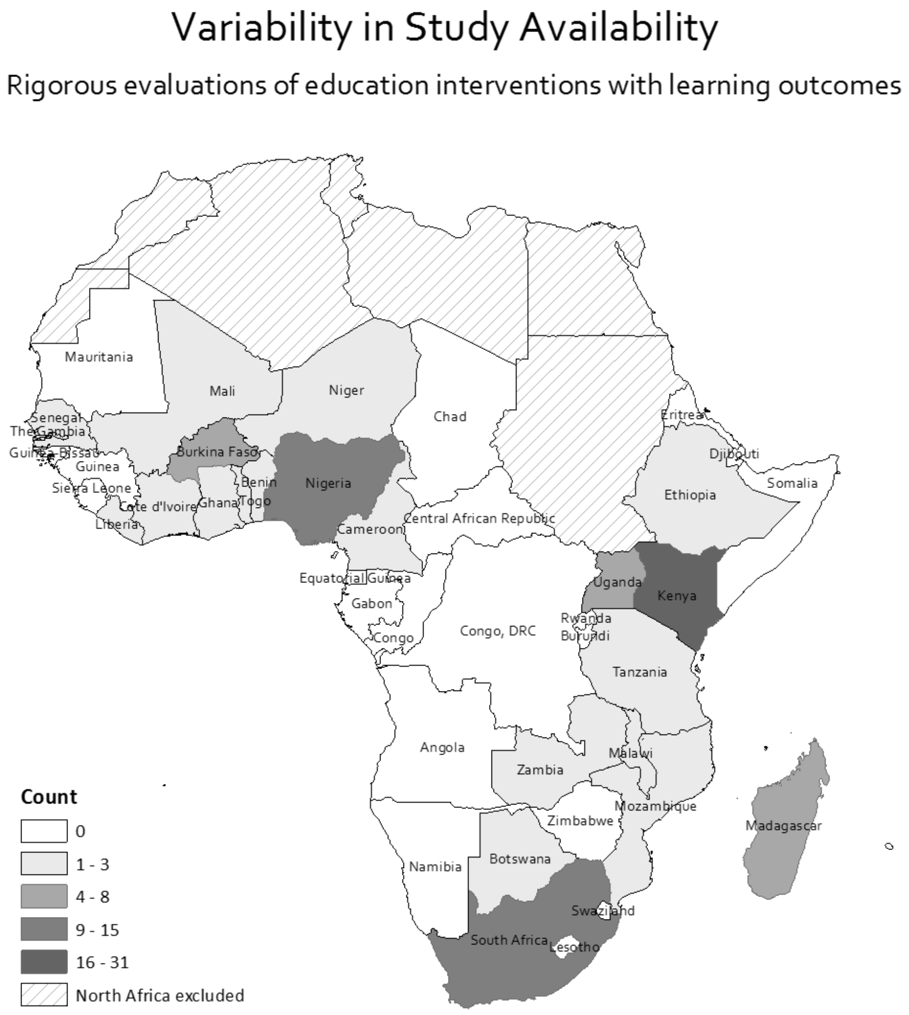
Regression Discontinuity	**0 points**	**1 point**	**2 points**	Total
Intervention/ Experiment Description	Description vague or unclear	Adequate description	Detailed description	2
Presentation	Very poor presentation/ poor language/ tables mis-labled	Adequate (poor language possible, but data tables are clear)	Professional quality presentation	2
Balance	Balance of groups (across discontinuity) not checked	Balance of groups (across discontinuity) checked (limited check on few variables)	Balance of groups (across discontinuity) fully checked (and any imbalance accounted for technically)	2
OVERALL				**6**

Matching	**0 points**	**1 point**	**2 points**	Total
Intervention/ Experiment Description	Description vague or unclear	Adequate description	Detailed description	2
Presentation	Very poor presentation/ poor language/ tables mis-labled	Adequate (poor language possible, but data tables are clear)	Professional quality presentation	2
Overlap	There is no check for overlap/ balance	There is a check for overlap/ balance, but any lack of overlap is not addressed	There is indeed overlap /lack of overlap is addressed	2
OVERALL				**6**

All indices are converted to a six points scale.

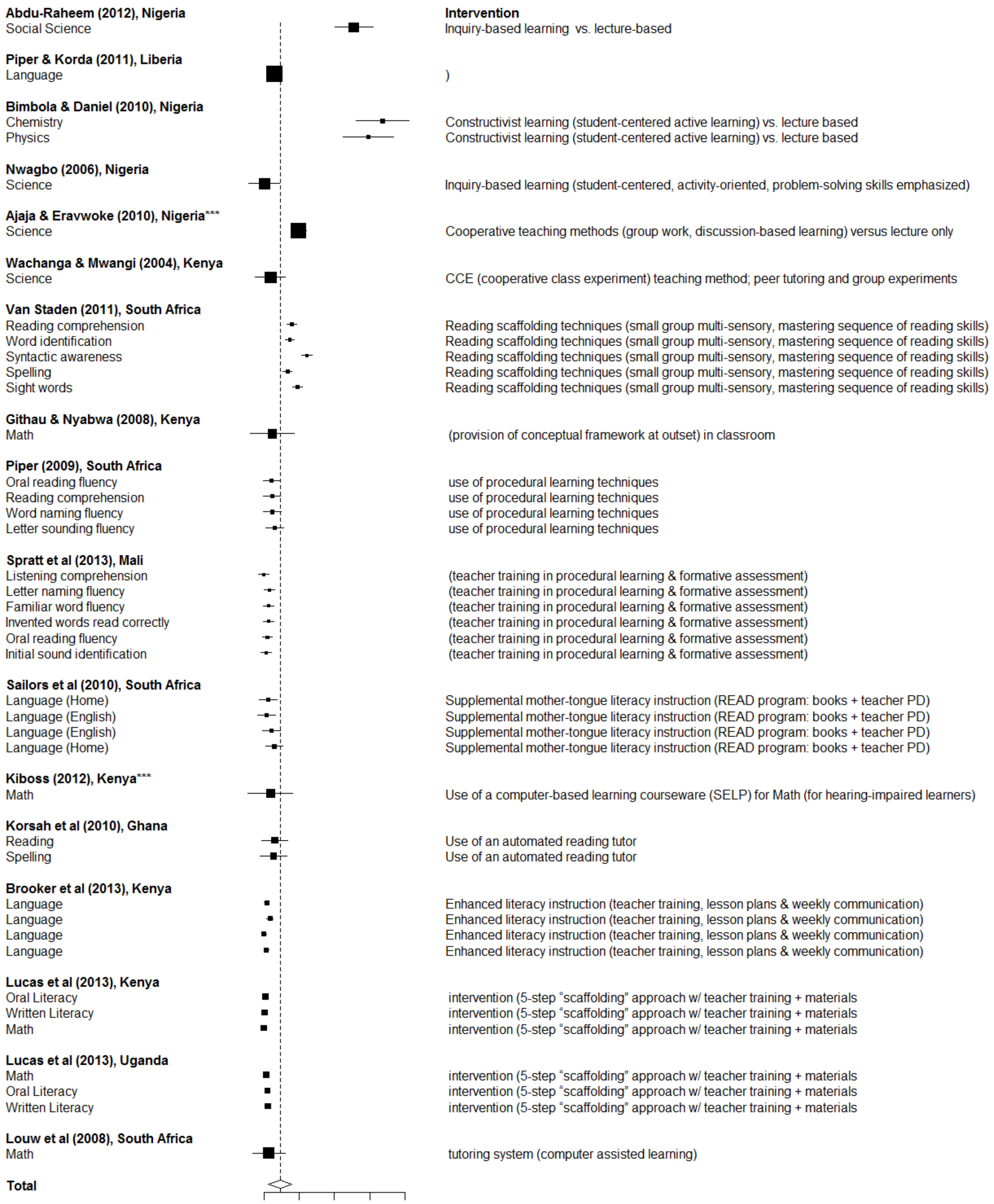
APPENDIX E.
Map of Study Locales

Figure 1 (Appendix E). Geographic Availability of Research



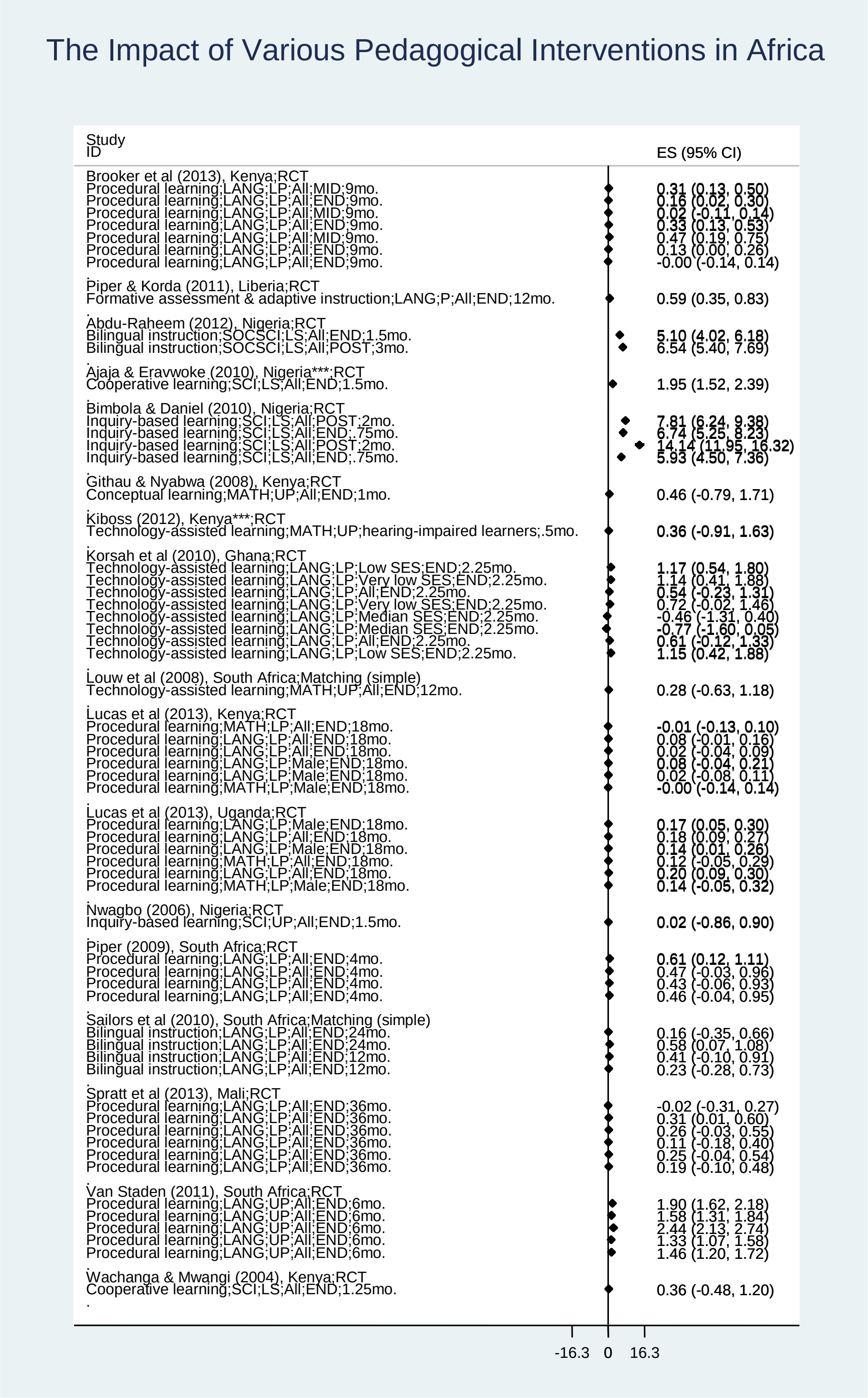
APPENDIX F.
Forest Plots & Effect Size Plots

Figure 2 (Appendix F). Forest Plot, Pedagogical Methods



NOTE: This is a "Forest Plot" and not an "Effect Size" plot. The "Total" pooled effect (above) is 0.181 standard deviation (ES = 0.045, df = 47.8, p = 0), as reported in the text. The effect size estimates are plotted around this pooled effect size and not 0 (as in the "Effect Size" plots which follow)

Figure 3 (Appendix F). Effect Size Plot, Pedagogical Methods



KEY:

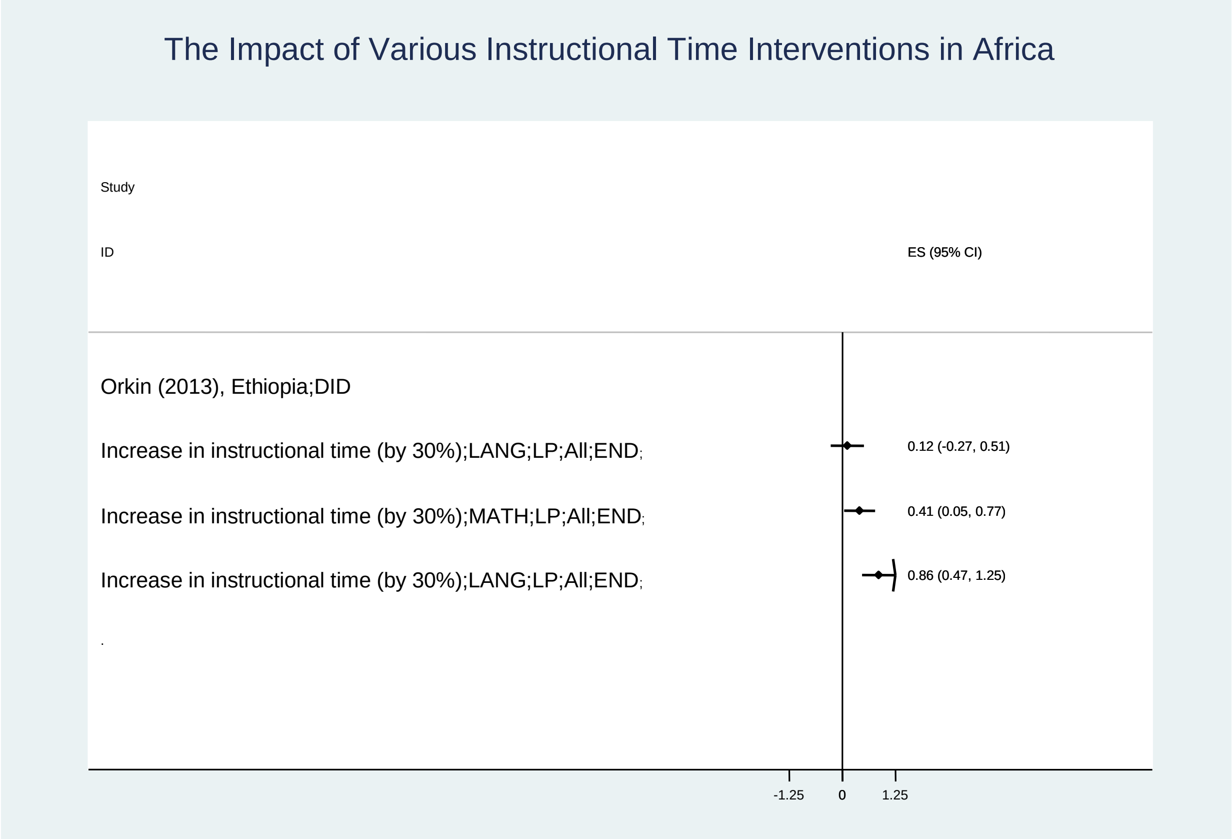
MATH=MATH **LANG** = LANGUAGE **COG**=COGNITION **COMP**=COMPOSITE

P = PRIMARY **S** = SECONDARY

END = AT END OF PROGRAM **POST** = POST END OF PROGRAM

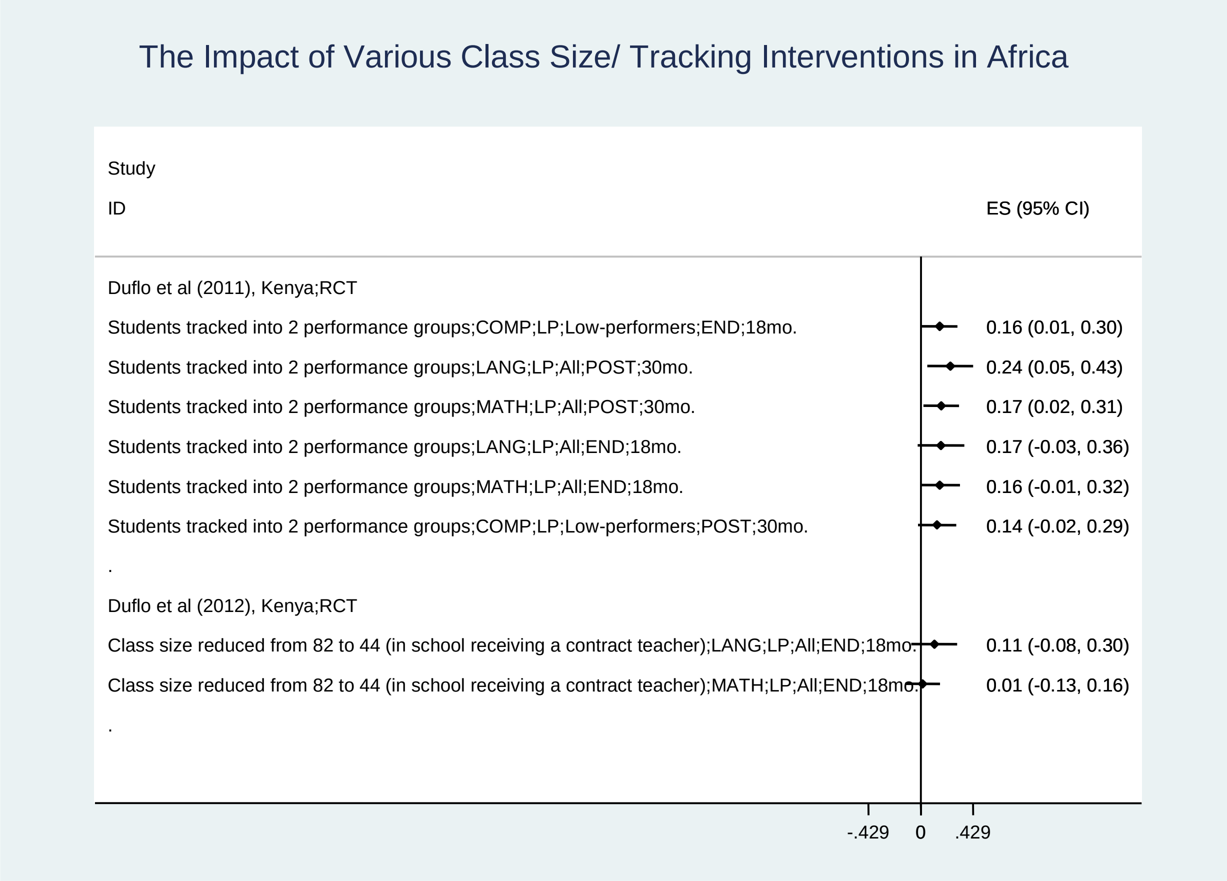
ALL = ALL STUDENTS **# MOS** = MONTHS

Figure 4 (Appendix F). Effect Size Plot, Instructional Time



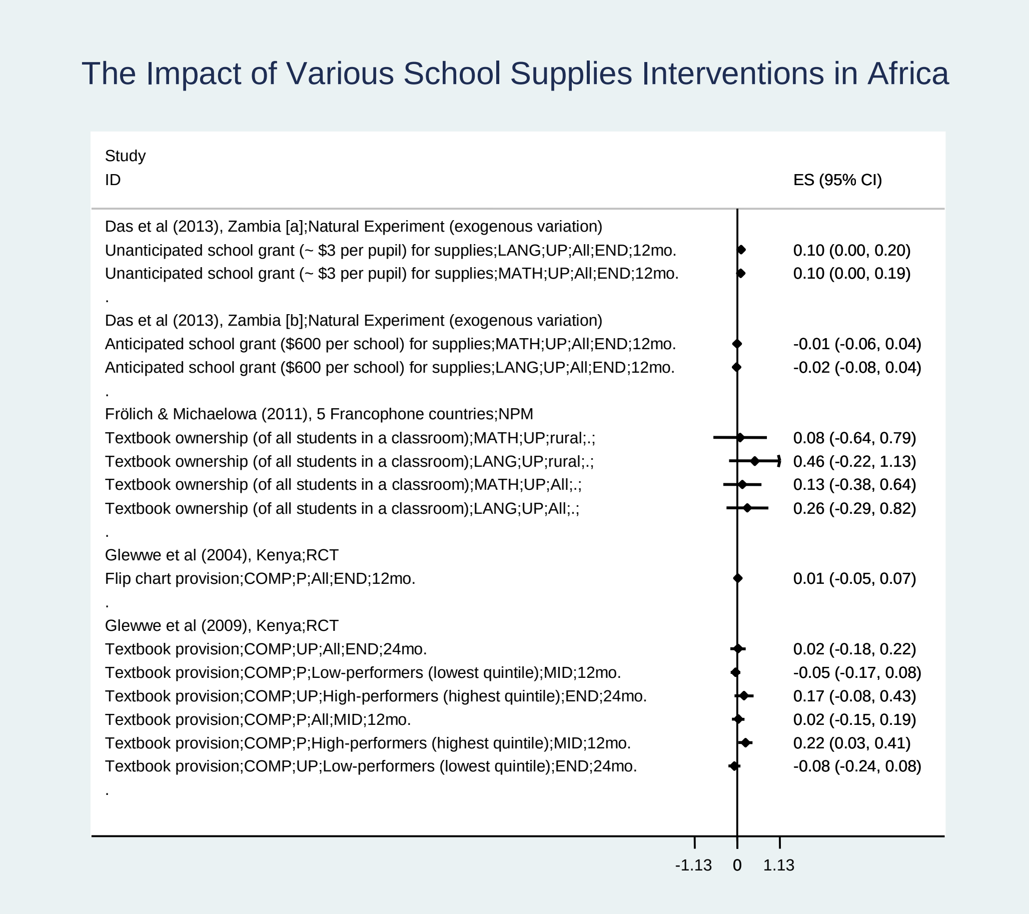
KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 5 (Appendix F). Effect Size Plot, Class Size



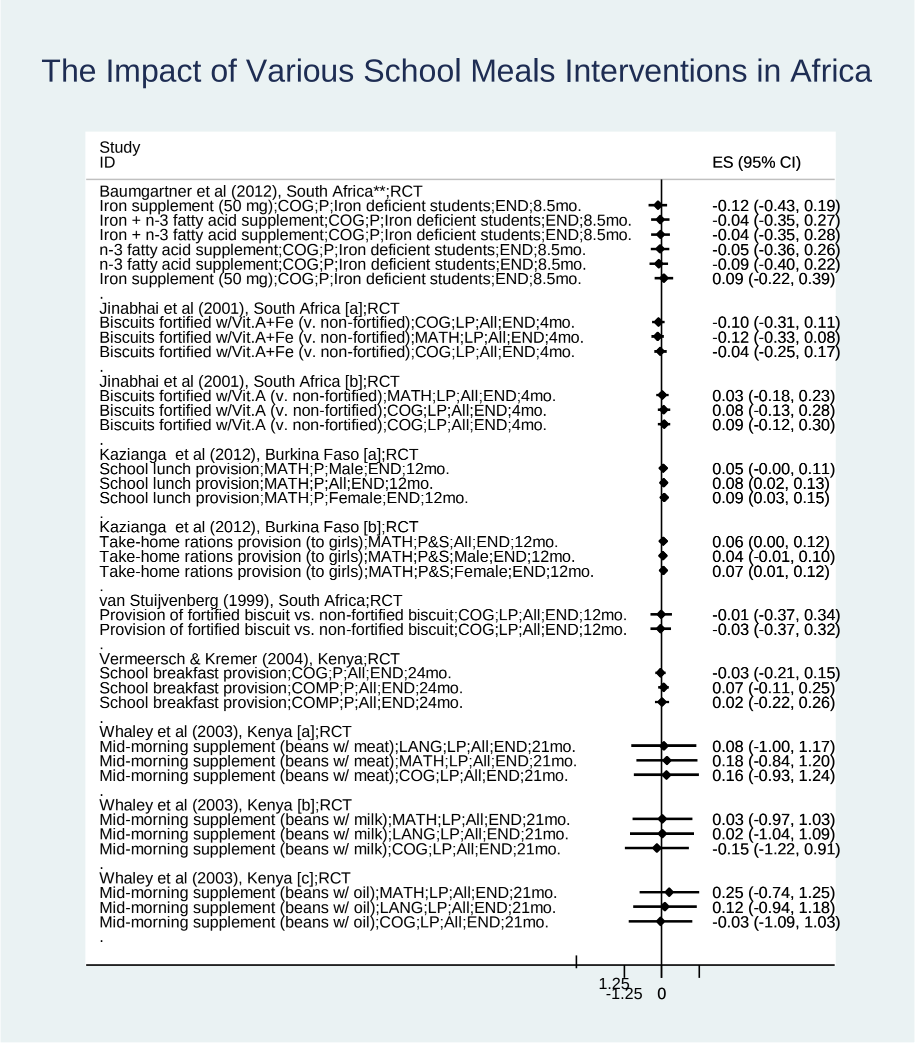
<u>KEY</u>:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 6 (Appendix F). Effect Size Plot, School Supplies



KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 7 (Appendix F). Effect Size Plot, School Meals



KEY:

MATH=MATH

LANG = LANGUAGE

COG=COGNITION

COMP=COMPOSITE

P = PRIMARY

S = SECONDARY

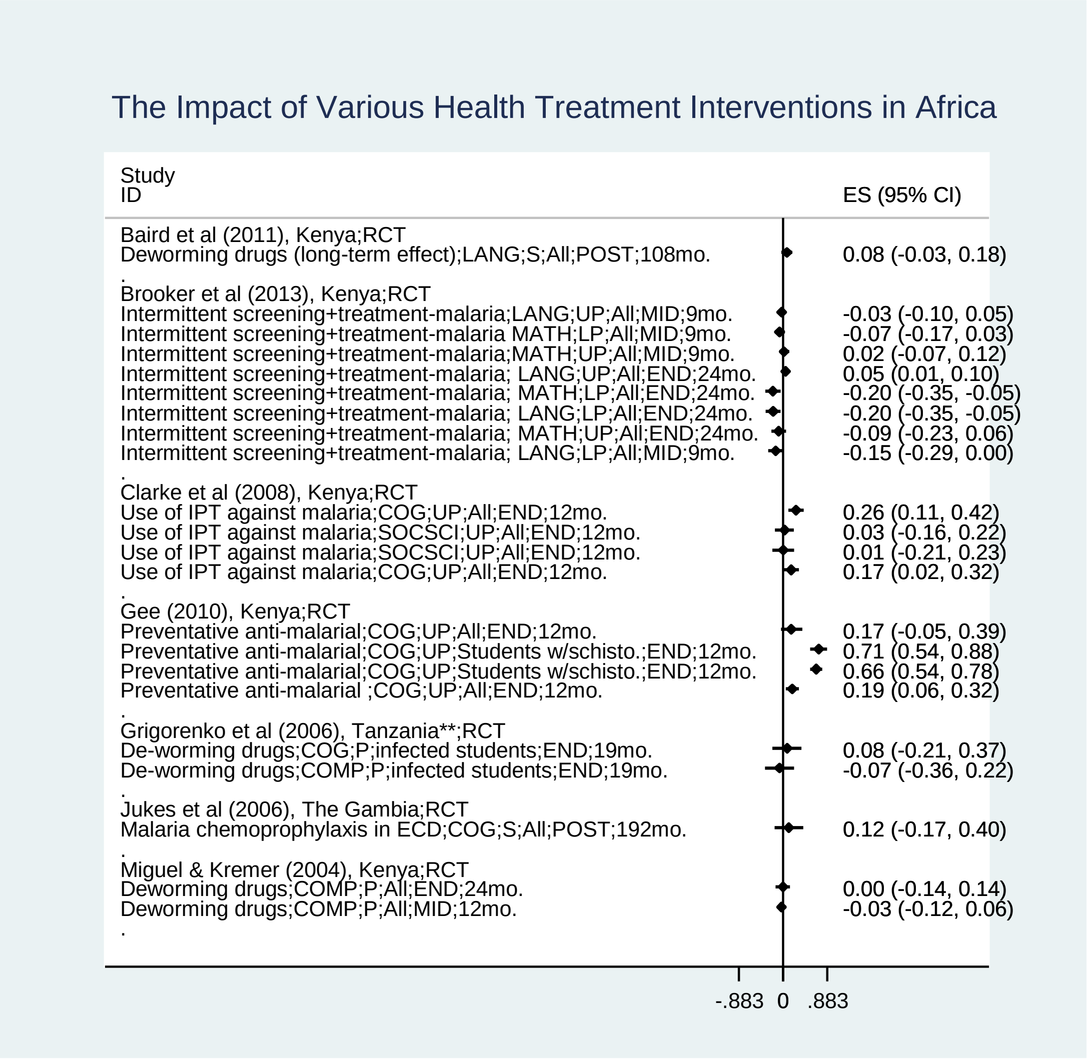
END = AT END OF PROGRAM

POST = POST END OF PROGRAM

ALL = ALL STUDENTS

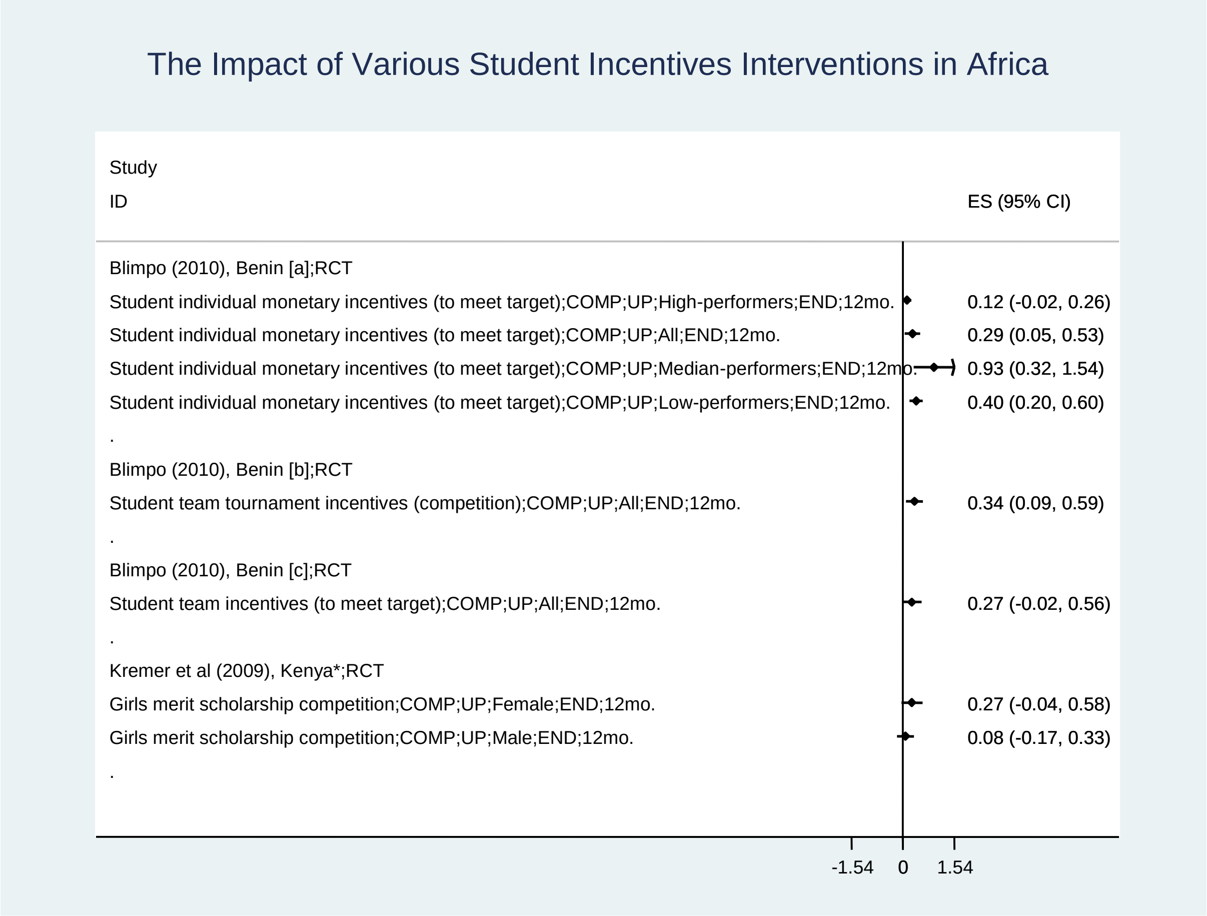
# MOS = MONTHS

Figure 8 (Appendix F). Effect Size Plot, Health Treatment



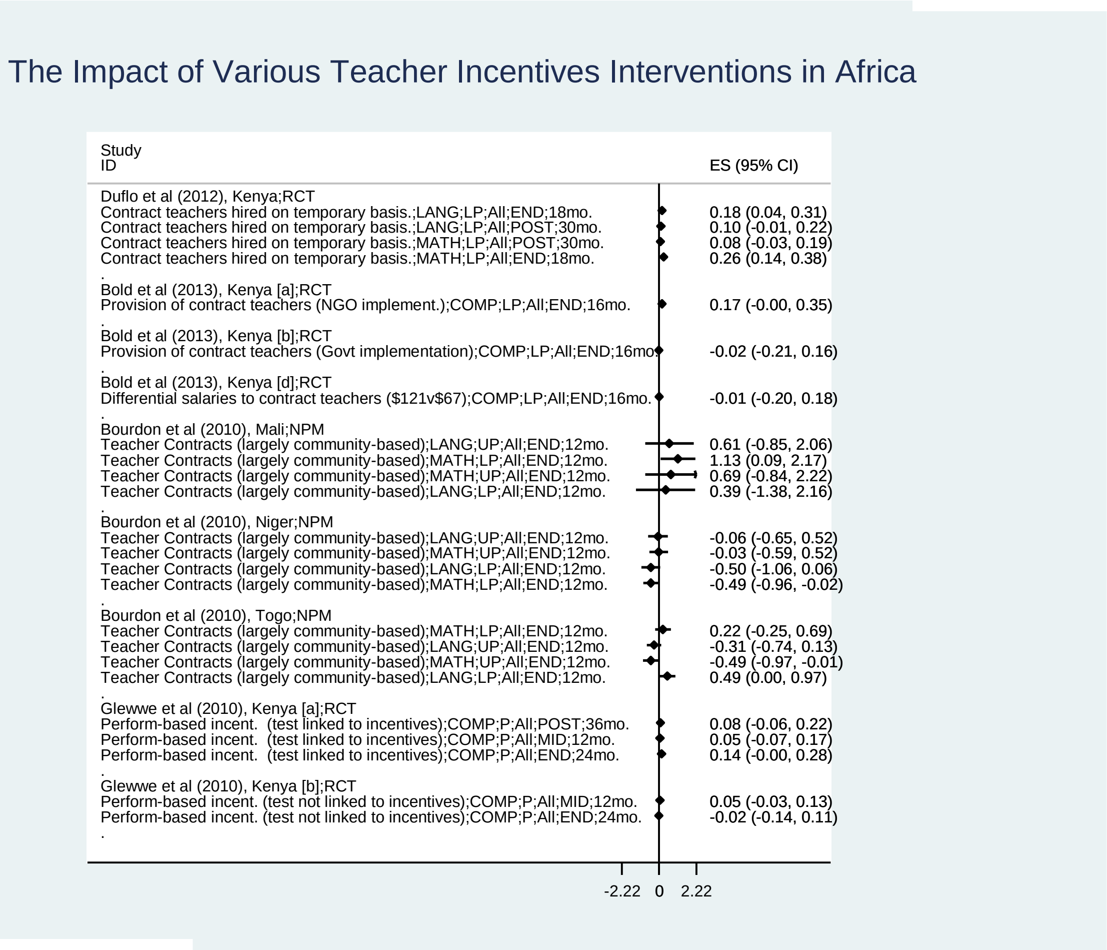
KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 9 (Appendix F). Effect Size Plot, Student Incentives



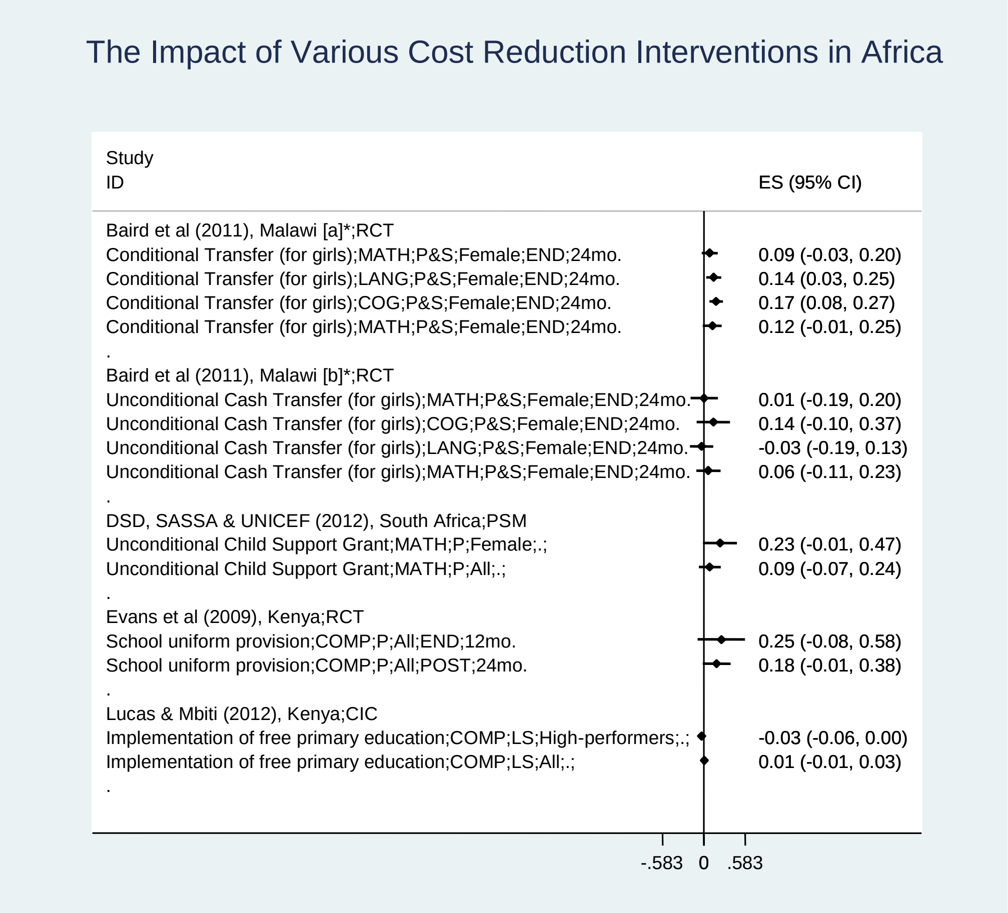
KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 10 (Appendix F). Effect Size Plot, Teacher Incentives



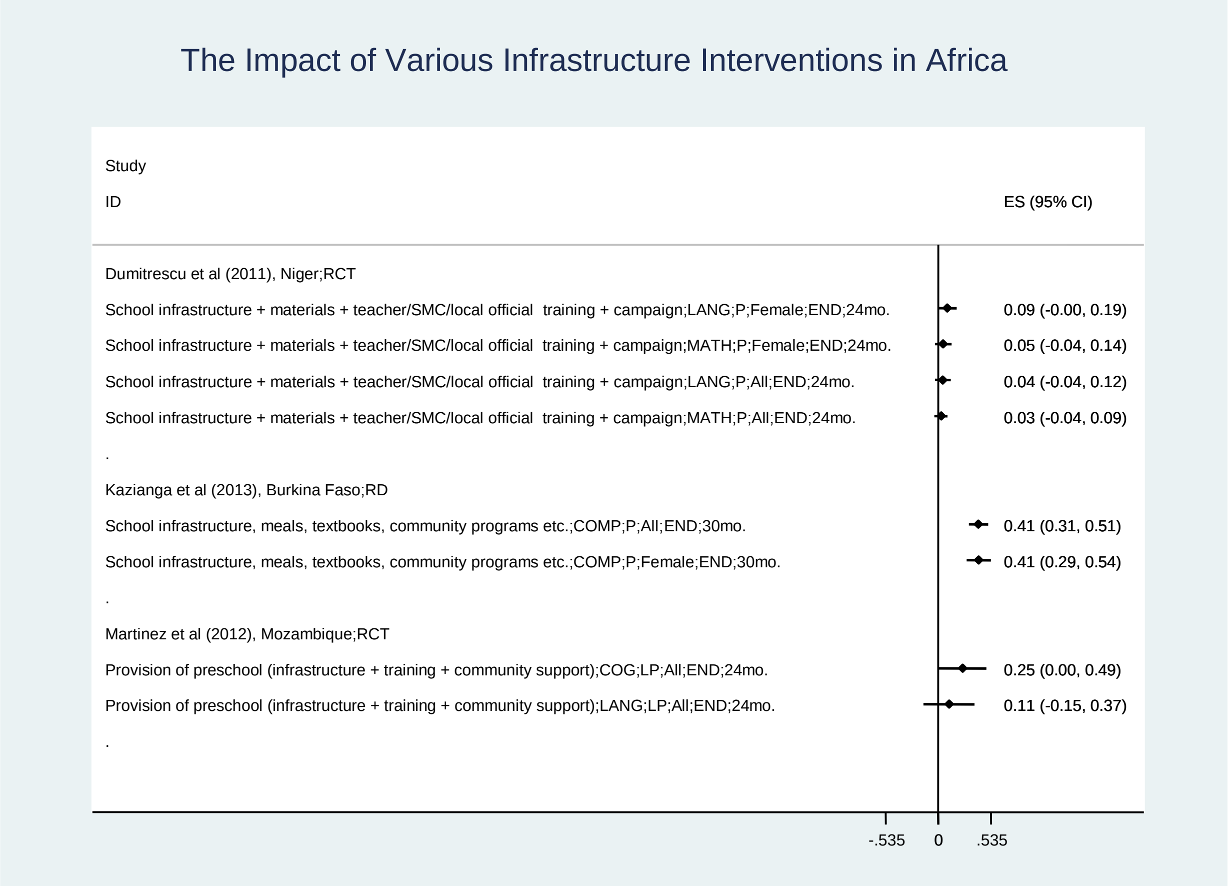
KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 11 (Appendix F). Effect Size Plot, Cost Reduction



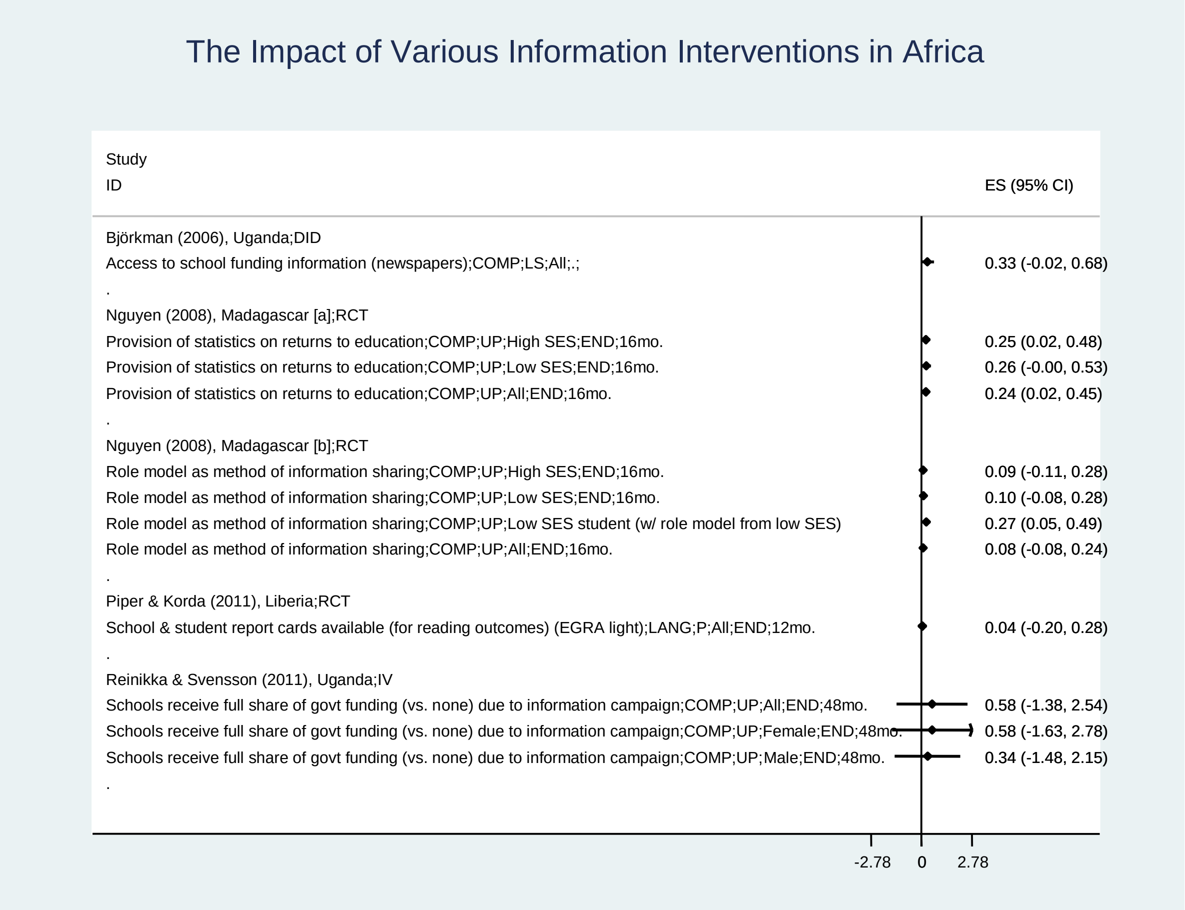
KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 12 (Appendix F). Effect Size Plot, Infrastructure & Complementary Inputs



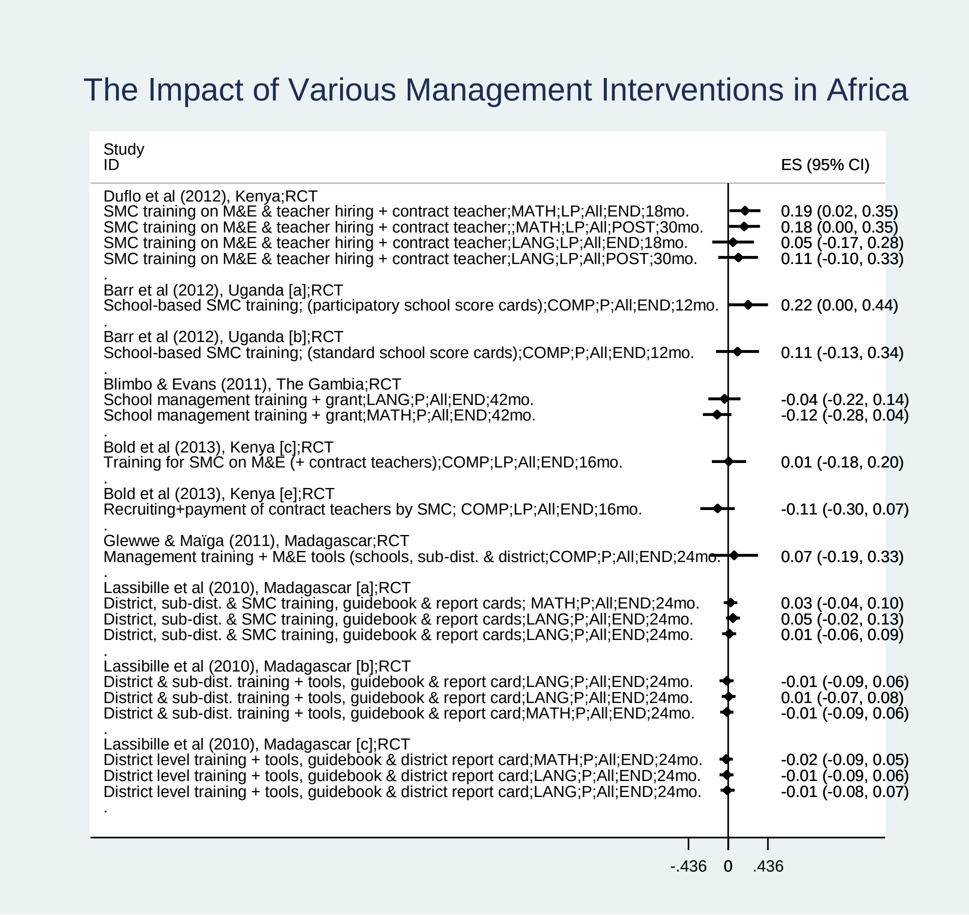
<u>KEY</u>:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 13 (Appendix F). Effect Size Plot, Information Interventions



KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

Figure 14 (Appendix F). Effect Size Plot, Management Interventions



<u>KEY</u>:

MATH=MATH

LANG = LANGUAGE

COG=COGNITION

COMP=COMPOSITE

P = PRIMARY

S = SECONDARY

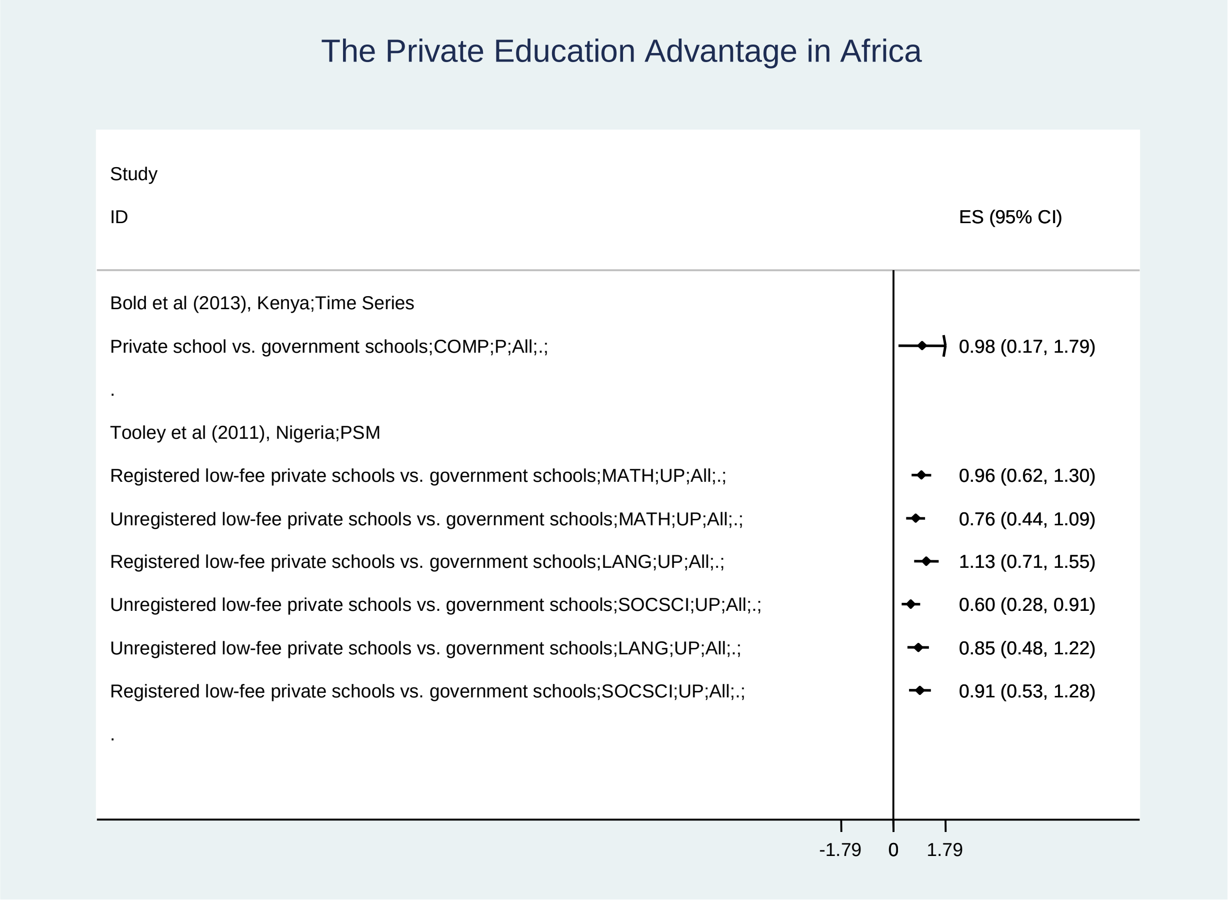
END = AT END OF PROGRAM

POST = POST END OF PROGRAM

ALL = ALL STUDENTS

# MOS = MONTHS

Figure 15 (Appendix F). Effect Size Plot, Private Education



KEY:
MATH=MATH
LANG = LANGUAGE
COG=COGNITION
COMP=COMPOSITE
P = PRIMARY
S = SECONDARY
END = AT END OF PROGRAM
POST = POST END OF PROGRAM
ALL = ALL STUDENTS
# MOS = MONTHS

APPENDIX G.
Additional Regression Results

Table 26 (Appendix G). Relative impact of pedagogical interventions (sample w/o outliers)

Sample= No outliers	Estimate	StdErr	df	P(|t|>)	95% CI.L	95% CI.U
Intercept	0.963*	0.454	8.64	0.064	-0.07	2.00
Pedagogical	0.257**	0.112	9.30	0.047	0.00	0.51
Assessment (researcher)	-0.025	0.078	14.37	0.756	-0.19	0.14
Assessment (standardized)	0.094	0.164	11.07	0.580	-0.27	0.46
Language	-0.044	0.071	17.17	0.550	-0.19	0.11
Math	-0.163	0.099	18.67	0.115	-0.37	0.04
Science	0.176	0.785	7.62	0.829	-1.65	2.00
Soc. Science	-0.214	0.275	8.46	0.457	-0.84	0.41
Reliability	0.091	0.268	8.36	0.742	-0.52	0.70
Quality	-0.093	0.072	9.15	0.229	-0.26	0.07
RCT	-0.153	0.160	4.25	0.391	-0.59	0.28
Matching	-0.265	0.186	7.36	0.195	-0.70	0.17
Primary	-0.036	0.118	5.00	0.773	-0.34	0.27
Work. Paper	0.012	0.067	14.86	0.866	-0.13	0.15
Report	-0.239	0.250	8.01	0.367	-0.82	0.34
East Afr.	-0.113	0.116	10.02	0.355	-0.37	0.15
Southern Afr.	-0.096	0.088	8.89	0.304	-0.29	0.10
Natl. Rep.	-0.126	0.077	9.52	0.134	-0.30	0.05
Length (mo.)	0.000	0.002	5.33	0.821	-0.01	0.00
Experiments	53.0					
I^2	80.530					
τ^2 estimate	0.043					

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used.
***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

Table 27 (Appendix G). Relative impact of school health interventions (full sample)

Full sample	Estimate	StdErr	df	P(|t|>)	95% CI.L	95% CI.U
Intercept	2.183*	1.127	9.1	0.085	-0.363	4.729
School Health	-0.449**	0.197	11.3	0.043	-0.881	-0.017
Assessment (researcher)	0.139	0.147	15.62	0.358	-0.173	0.451
Assessment (standardized)	0.580*	0.320	11.55	0.096	-0.120	1.279
Language	-0.048	0.095	16.86	0.622	-0.248	0.152
Math	-0.119	0.119	18.93	0.331	-0.368	0.130
Science	-0.157	0.966	9.48	0.874	-2.326	2.012
Soc. Science	0.586	1.303	2.76	0.686	-3.774	4.946
Reliability	0.012	0.352	9.99	0.973	-0.772	0.796
Quality	-0.239	0.146	9.94	0.132	-0.565	0.086
RCT	-0.332	0.226	4.09	0.215	-0.954	0.291
Matching	-0.209	0.268	6.68	0.461	-0.848	0.429
Primary	-0.361	0.385	6.14	0.383	-1.297	0.575
Work. Paper	0.032	0.108	15.42	0.768	-0.197	0.262
Report	-0.164	0.210	8.23	0.456	-0.646	0.317
East Afr.	-0.020	0.146	11.02	0.895	-0.340	0.301
Southern Afr.	-0.459	0.260	9.68	0.109	-1.041	0.123
Natl. Rep.	-0.255*	0.117	10.94	0.052	-0.513	0.002
Length (mo.)	-0.001	0.005	6.41	0.899	-0.012	0.011

Number of studies = 55

Number of outcomes = 130 (min = 1 , mean = 2.36 , median = 2 , max = 9)

ρ = 0.8

I^2 = 84.03

τ^2 estimate= 0.058

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used.
***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

Table 28 (Appendix G). Relative impact of school health programs (high quality sample)

Sample= High quality only	Estimate	StdErr	df	P(|t|>)	95% CI.L	95% CI.U
Intercept	0.596*	0.231	4	0.056	-0.023	1.215
School Health	-0.119*	0.057	10	0.064	-0.246	0.008
Assessment (researcher)	-0.007	0.087	11	0.939	-0.200	0.186
Assessment (standardized)	0.222	0.175	10	0.231	-0.165	0.610
Math	-0.058	0.045	19	0.214	-0.152	0.036
Language	-0.012	0.043	16	0.785	-0.103	0.079
Reliability	0.065	0.087	8	0.478	-0.136	0.265
RCT	-0.232	0.149	5	0.183	-0.621	0.157
Matching	-0.117	0.237	5	0.643	-0.727	0.493
Primary	-0.200	0.141	4	0.228	-0.590	0.190
Work. Paper	0.055	0.038	14	0.170	-0.026	0.136
Report	-0.008	0.076	7	0.919	-0.188	0.172
East Afr.	-0.033	0.092	7	0.726	-0.248	0.181
Southern Afr.	-0.215*	0.108	6	0.094	-0.479	0.050
Natl. Rep.	-0.132**	0.054	10	0.035	-0.253	-0.012
Length (mo.)	-0.002	0.002	6	0.417	-0.007	0.003

Number of studies = 47

Number of outcomes = 114 (min = 1 , mean = 2.43 , median = 2 , max = 9)

ρ = 0.8

I^2 = 50.502

τ^2 estimate= 0.010

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used.
***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

Table 29 (Appendix G). Explaining heterogeneity in full & high quality samples

Sample		Estimate	Std. Err.	df	P(|t|>)	95% CI.L	95% CI.U	*I2*	*Studies*
Full	Intercept	0.175**	0.065	8.6	0.03	0.03	0.32	*89.3*	*60*
	RCT	0.023	0.086	12.5	0.80	-0.16	0.21		
High Quality	Intercept	0.156*	0.073	5.59	0.08	-0.03	0.34	*68.5*	*52*
	RCT	-0.077	0.076	8.15	0.34	-0.25	0.10		
Full	Intercept	0.187***	0.048	43.65	0.00	0.09	0.28	*89.5*	*60*
	Matching	-0.087	0.086	4.55	0.36	-0.31	0.14		
High Quality	Intercept	0.089***	0.021	34.67	0.00	0.05	0.13	*69.2*	*52*
	Matching	0.003	0.054	2.74	0.96	-0.18	0.18		
Full	Intercept	0.178**	0.061	9.13	0.02	0.04	0.32	*89.3*	*60*
	ITT est.	0.018	0.084	13.57	0.83	-0.16	0.20		
High Quality	Intercept	0.160*	0.070	6.11	0.06	-0.01	0.33	*68.6*	*52*
	ITT est.	-0.083	0.072	9.09	0.28	-0.25	0.08		
Full	Intercept	0.208***	0.053	44.3	0.00	0.10	0.31	*88.8*	*60*
	ATT est.	-0.139	0.081	5.35	0.14	-0.34	0.06		
High Quality	Intercept	0.095***	0.021	35.31	0.00	0.05	0.14	*65.0*	*52*
	ATT est.	-0.045	0.046	2.91	0.40	-0.19	0.10		
Full	Intercept	0.282**	0.108	20.7	0.02	0.06	0.51	*89.5*	*60*
	Work.	-0.169	0.111	39.2	0.14	-0.39	0.06		
	Report	-0.135	0.128	14.6	0.31	-0.41	0.14		
High Quality	Intercept	0.073*	0.035	13.3	0.05	0.00	0.15	*67.5*	*52*
	Work.	0.027	0.042	29.3	0.53	-0.06	0.11		
	Report	0.032	0.072	11.7	0.67	-0.13	0.19		
Full	Intercept	0.099***	0.020	36.2	0.00	0.06	0.14	*88.6*	*60*
	Reliability	0.398*	0.203	16.6	0.07	-0.03	0.83		
High Quality	Intercept	0.088***	0.019	30.84	0.00	0.05	0.13	*69.1*	*52*
	Reliability	0.014	0.085	7.78	0.87	-0.18	0.21		
Full	Intercept	0.172***	0.054	12.5	0.01	0.05	0.29	*89.0*	*60*
	Assess.(r)†	0.024	0.081	20.7	0.77	-0.14	0.19		
High Quality	Intercept	0.132**	0.053	8.4	0.04	0.01	0.25	*66.4*	*52*
	Assess.(r)†	-0.052	0.057	13.5	0.38	-0.18	0.07		
Full	Intercept	0.232***	0.061	36.7	0.00	0.11	0.36	*88.9*	*59*
	Natl. rep.	-0.177**	0.069	17.5	0.02	-0.32	-0.03		
High Quality	Intercept	0.104***	0.024	28.2	0.00	0.05	0.15	*64.8*	*51*
	Natl. rep	-0.060	0.035	13.9	0.11	-0.14	0.02		
Full	Intercept	0.328	0.193	13.75	0.11	-0.09	0.742	*91.9*	*27*
	SJR rank	-0.026	0.021	4.54	0.27	-0.08	0.029		
High Quality	Intercept	0.053	0.040	8.71	0.22	-0.04	0.144	*74.5*	*23*
	SJR rank	0.003	0.005	3.94	0.62	-0.01	0.018		

[TABLE CONTINUED ON FOLLOWING PAGE]

[Table 29 continued].

Sample		Estimate	Std. Err.	df	P(|t|>)	95% CI.L	95% CI.U	*I2*	*Studies*
Full	Intercept	0.059	0.049	9.74	0.25	-0.05	0.168	*73.9*	*23*
	AI rank	0.002	0.005	3.35	0.75	-0.01	0.017		
High Quality	Intercept	0.056	0.049	9.63	0.28	-0.05	0.165	*75.0*	*22*
	AI rank	0.002	0.005	3.34	0.70	-0.01	0.017		
Full	Intercept	0.375**	0.146	11.4	0.03	0.06	0.694	*88.9*	*60*
	East	-0.284*	0.147	21.1	0.07	-0.59	0.022		
	Southern	-0.139	0.202	21.1	0.50	-0.56	0.281		
High Quality	Intercept	0.161*	0.072	6..9	0.06	-0.01	0.33	*65.1*	*52*
	East	-0.088	0.074	12.2	0.26	-0.25	0.07		
	Southern	-0.097	0.075	13.5	0.22	-0.26	0.06		
Full	Intercept	0.072***	0.222	21.8	0.00	0.03	0.118	*88.6*	*60*
	Supply	0.216*	0.086	46.6	0.02	0.04	0.389		
High Quality	Intercept	0.057***	0.019	18.4	0.00	0.02	0.097	*64.8*	*52*
	Supply	0.065*	0.038	36.0	0.10	-0.01	0.14		

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

†The assessment was designed by researcher.

Table 30 (Appendix G). Differential Impact of short vs. long-term school meal programs

	Estimate	std. err.	t-val	df	P(|t|>)	95% CI.L	95% CI.U
Intercept	-0.027*	0.005	-5.726	1.08	‡	-0.078	0.024
Less than 9 mos.	0.020	0.023	0.844	2.04	‡	-0.079	0.119

Number of studies = 5

Number of outcomes = 16 (min = 1 , mean = 3.2 , median = 3 , max = 6)

ρ = 0.8

I^2 = 0

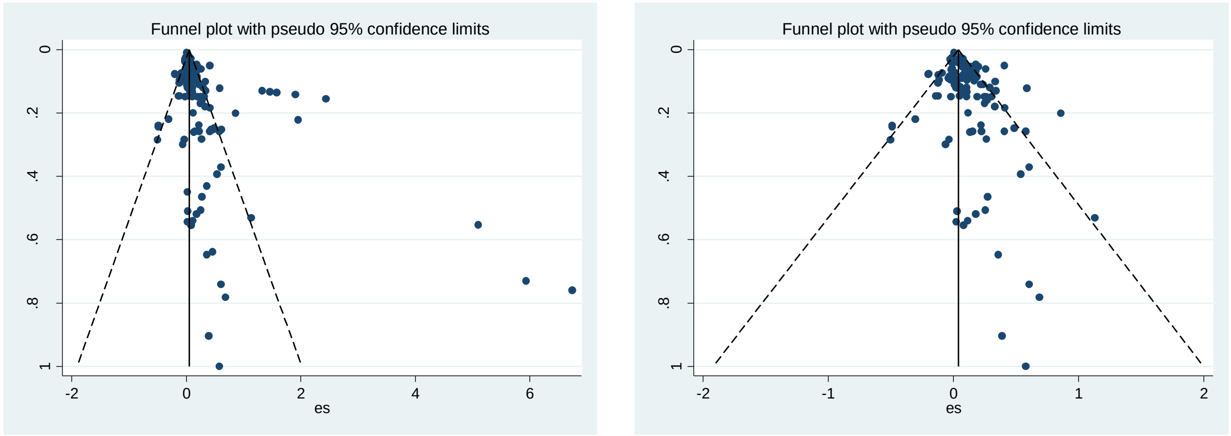
τ^2 estimate= 0

Robust Variance Estimation is used to cluster standard errors within experiments. Small sample corrections are used. ***, **, * indicate statistical significance at 1%, 5%, and 10%, respectively.

‡ When the degrees of freedom are less than 4, Tipton (in press) notes that the normal approximation fails and p-values should not be interpreted

APPENDIX H.
Publication Bias Plot Series

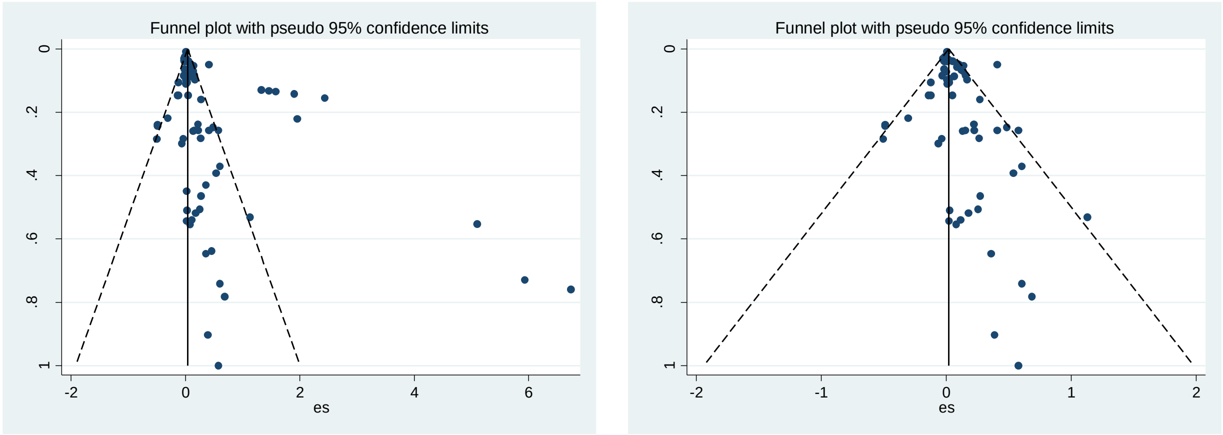
Figure 16 (Appendix H). Plot Series: Publication bias, full sample & high quality sample



Full Sample | High Quality Sample

Egger's test: bias coeff.: 1.84, p-val= 0.00 | *Egger's test: bias coeff.: 0.795, p-val= 0.00*

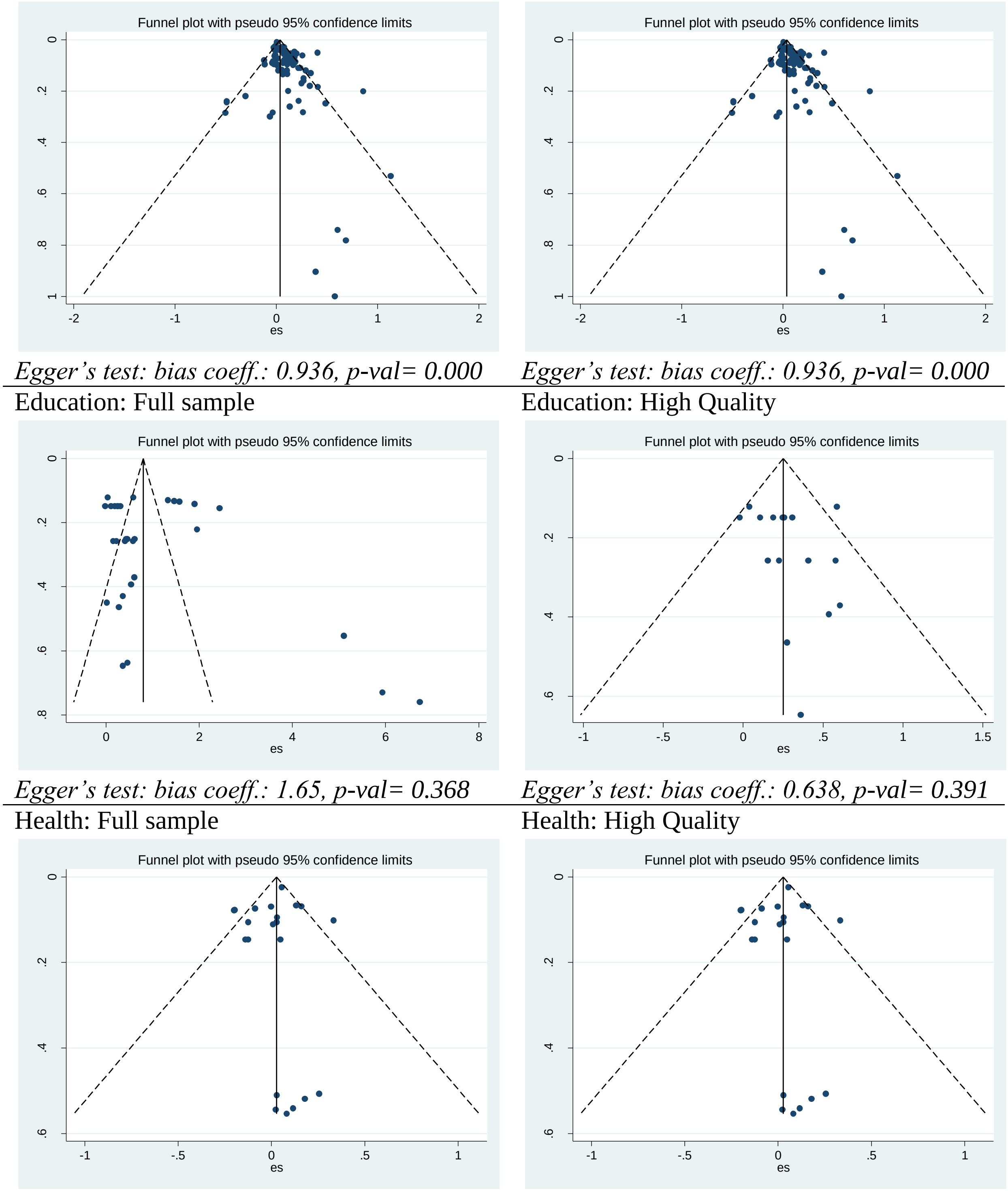
Figure 17 (Appendix H). Plot Series: Publication bias, sample w/ only journal articles



Journal Articles (full sample) | Journal Articles (high quality)

Egger's test: bias coeff.: 2.11, p-val= 0.00 | *Egger's test: bias coeff.: 0.527, p-val= 0.026*

Figure 18 (Appendix H). Plot Series: Publication bias by academic field



Economics: Full sample

Egger's test: bias coeff.: 0.936, p-val= 0.000

Economics: High Quality

Egger's test: bias coeff.: 0.936, p-val= 0.000

Education: Full sample

Egger's test: bias coeff.: 1.65, p-val= 0.368

Education: High Quality

Egger's test: bias coeff.: 0.638, p-val= 0.391

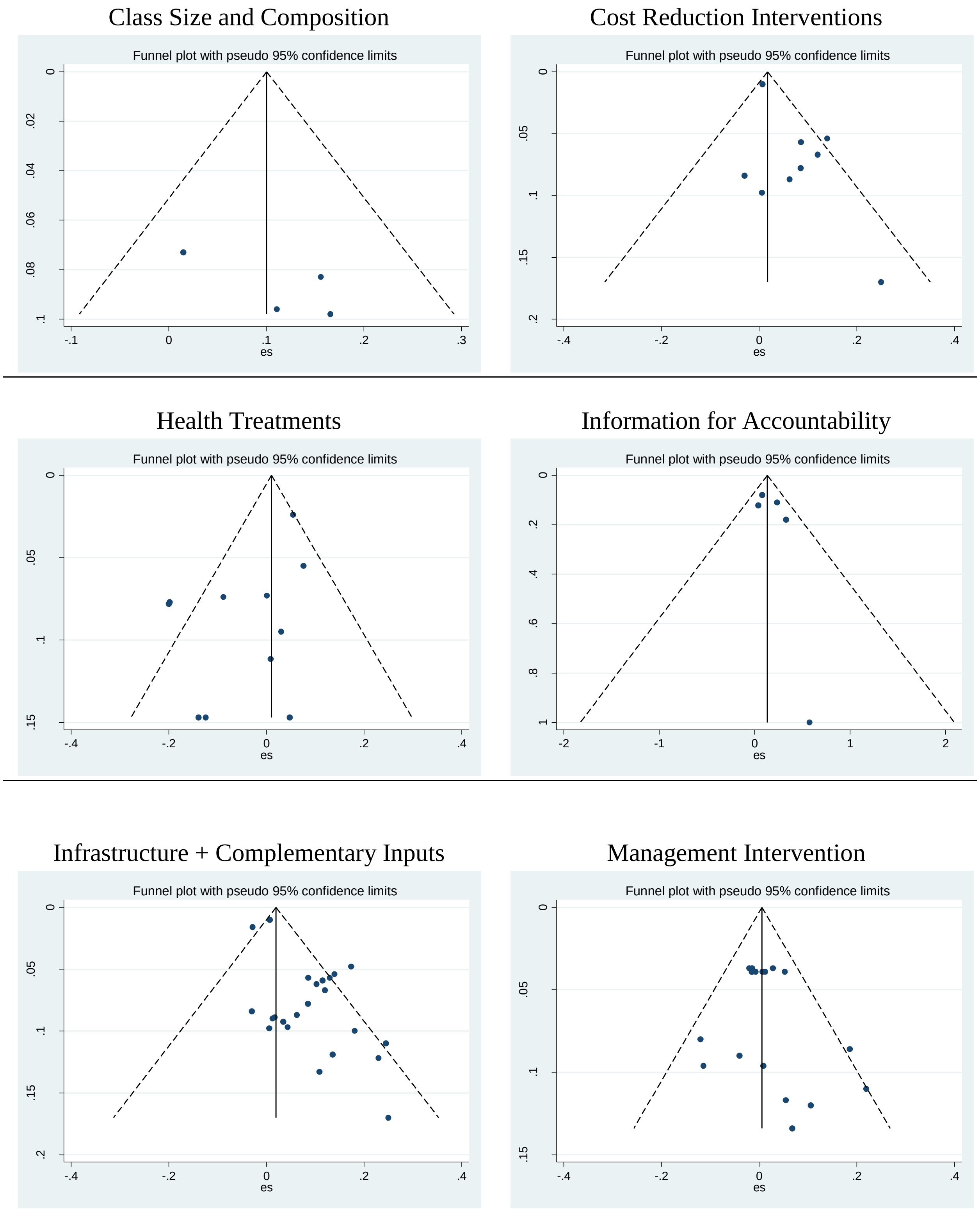
Health: Full sample

Egger's test: bias coeff.: -1.53, p-val= 0.056

Health: High Quality

Egger's test: bias coeff.: -1.53, p-val= 0.056

Figure 19 (Appendix H). Plot Series: Publication bias by topic



Class Size and Composition

Cost Reduction Interventions

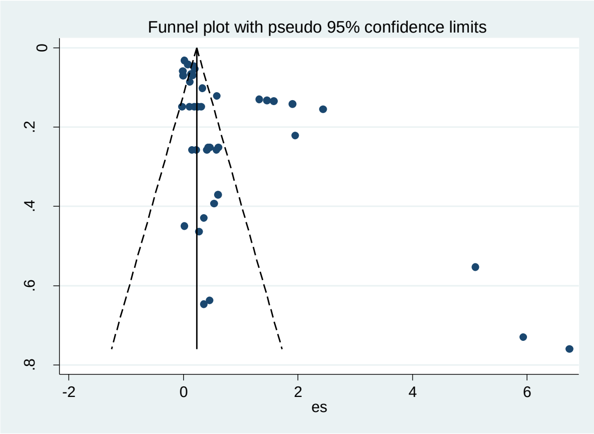
Health Treatments

Information for Accountability

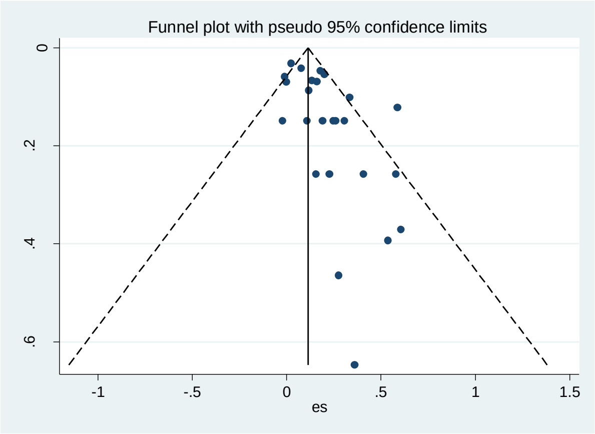
Infrastructure + Complementary Inputs

Management Intervention

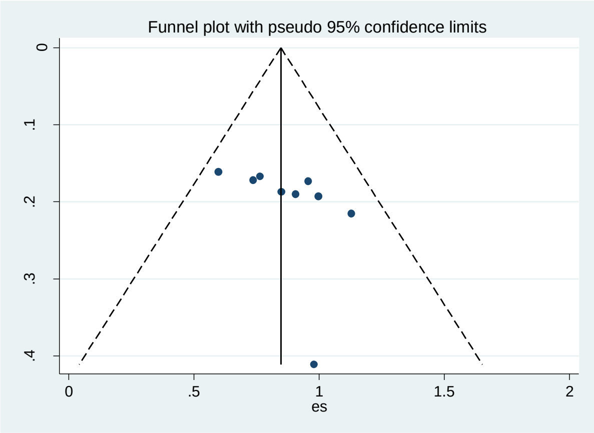
Pedagogical Methods



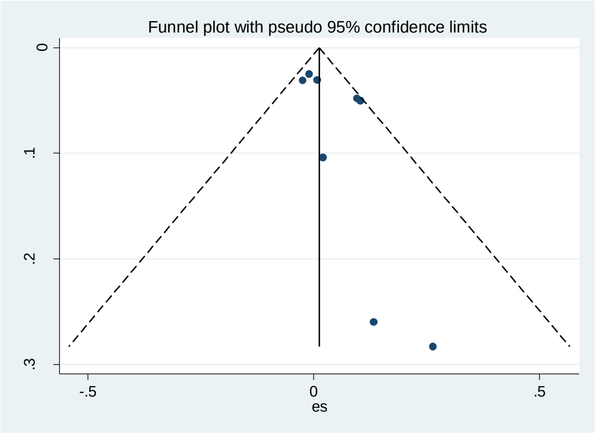
Pedagogical Methods (high quality)



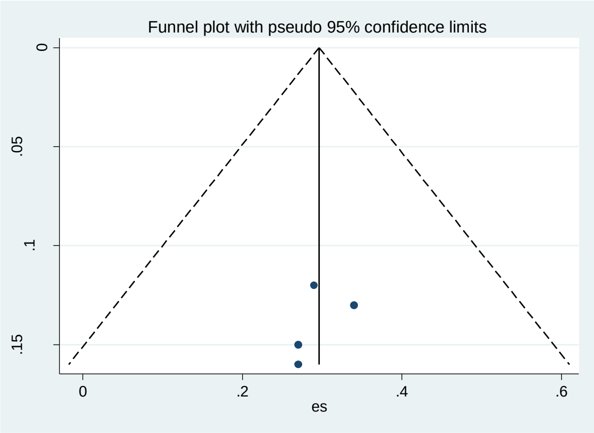
Private Education



School Supplies Intervention



Student Incentives



Teacher Incentives

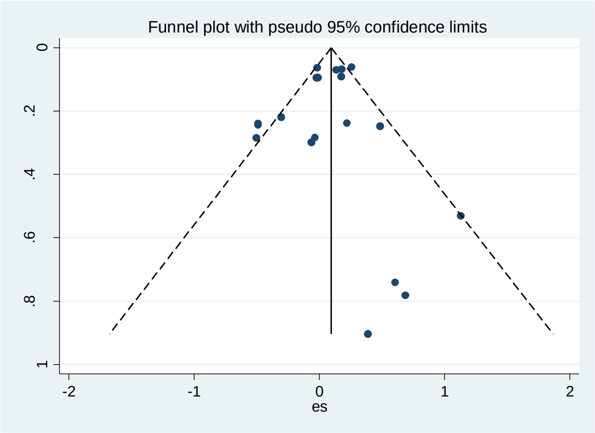
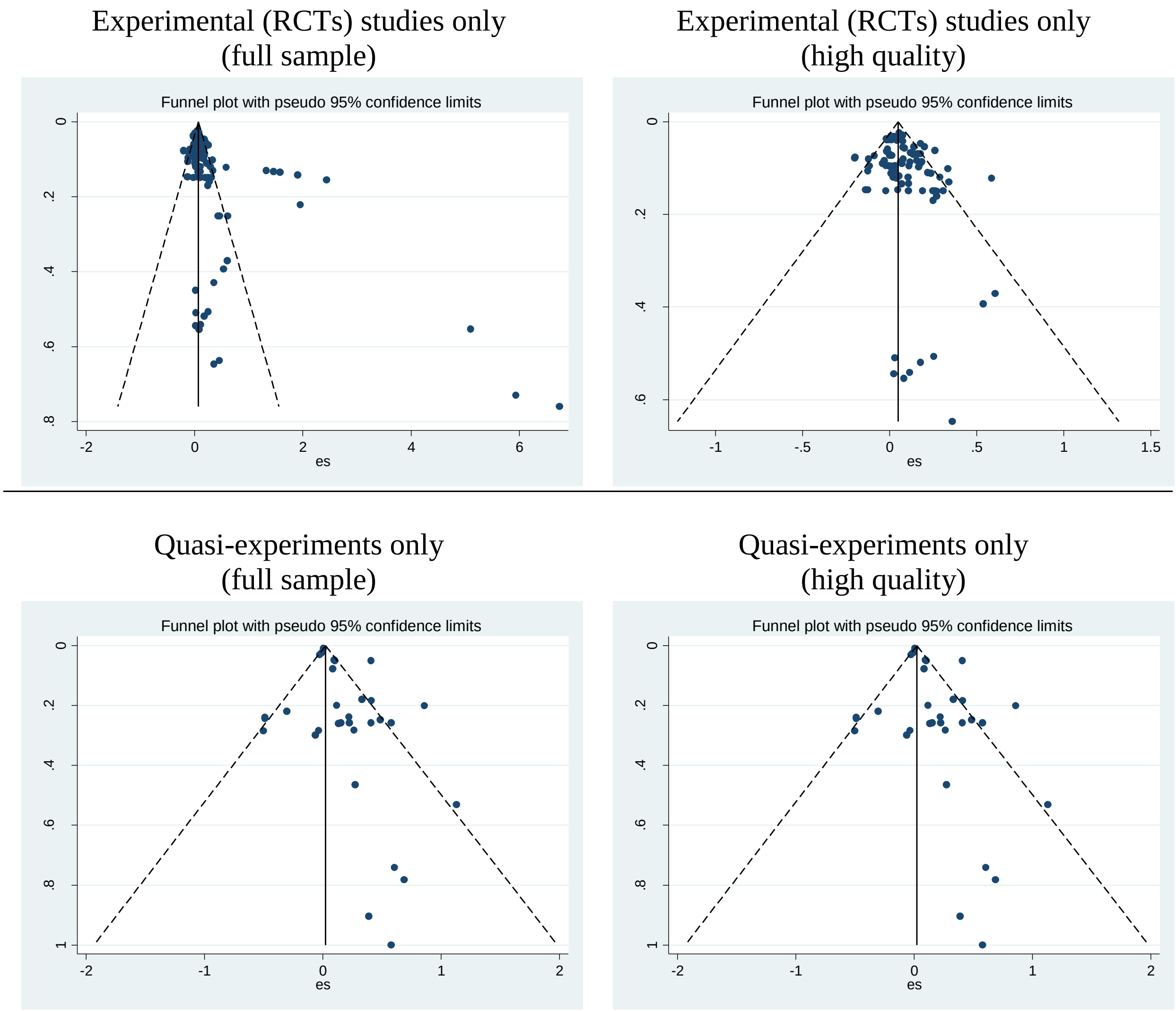


Figure 20 (Appendix H). Plot Series: Publication bias by identification strategy


Experimental (RCTs) studies only (full sample)

Experimental (RCTs) studies only (high quality)

Quasi-experiments only (full sample)

Quasi-experiments only (high quality)

APPENDIX I.
Study Coding Sheet (Round 1)

SECTION A. ENTRY DESCRIPTIVES		
A1.	Entry person name *(last, first)*	______________________,_______________________
A2.	Assignment of study ID *(3 more digits, starting with*	| R | 0 | 0 |__|__|__|
A3.	Entry round of study *(specify 1-6)*	____
A4.	Entry date of study	|__|__|/ |__|__|/|__|__|__|__| month/ day/ year
A5.	What is the intervention area?	1. [ ] Abolishment of School Fees 2. [ ] Charter Schools 3. [ ] Class Size 4. [ ] Comprehensive Equity Program 5. [ ] Conditional Cash Transfers 6. [ ] Information for Accountability 7. [ ] Private Education 8. [ ] School Health 9. [ ] School Resource Provision 10. [ ] School-Based Management 11. [ ] Student Scholarships & Incentives 12. [ ] Teacher Incentives 13. [ ] Tracking & Peer Effects 14. [ ] *Decentralization* 15. [ ] High-Stakes Testing & Accountability Systems 16. [ ] Information and Communications Technology 17. [ ] Instructional Time 18. [ ] School Choice 19. [ ] School Infrastructure 20. [ ] Supplemental Instruction/ Tutoring 21. [ ] Vouchers *22.* [ ] Other (describe) ____
A6.	Is this study entered in duplicate?	1. [ ] Yes 2. [ ] No >> SKIP to #A5.
A7.	IF YES, what is the reason for the duplication?	____
A8.	Comments on status of entry	____ ____
A9.	Has the effect size entry been checked for reliability	1. [ ] Yes 2. [ ] No

SECTION B. BIOGRAPHICAL INFORMATION

B0.	What is the full academic citation of this study?	____ ____
B1.	What are the authors last names, year and location?	________________________(_____), _____________ Authors last names (year), location *If date is before 1980 >> EXCLUDE ◈
B2.	What are the authors' institutions?	_____________ (________________________________) last name (institution) _____________ (________________________________) last name (institution) _____________ (________________________________) last name (institution) _____________ (________________________________) last name (institution) _____________ (________________________________) last name (institution)
B3.	What is the study publication type?	1. [ ] Journal Article 2. [ ] Working Paper 3. [ ] Report 4. [ ] Book/Book Chapter 5. [ ] Conference Presentation 6. [ ] Dissertation 7. [ ] Draft Paper 8. [ ] Other (describe ____________________)
B4.	Does the study originate from any of the following research centers?	1. [ ] J-PAL 2. [ ] NBER 3. [ ] Mathematica 4. [ ] World Bank 5. [ ] Other independent research center (describe _____________________________)
B5.	Notes on publication type	____ ____

SECTION C. INTERVENTION/ DESIGN

C1.	In what region of the world does the study take place?	1. [ ] Sub-Saharan Africa 2. [ ] Middle East & North Africa 3. [ ] Other *If other >> EXCLUDE ◈
C2.	In what country (within Africa) does the study take place?	
C3.	What is the setting of the study?	1. [ ] Rural 2. [ ] Urban 3. [ ] National-representative 4. [ ] Other (describe:
C4.	Abstract of article (if no abstract exists, use introductory paragraph) *(may be easiest to simply cut and paste into database directly).*	____________________________________ __ ____________________________________ __ ____________________________________ __
C5.	Any additional intervention-specific details	____________________________________ __
C6.	How many treatment groups are in the study?	1. [ ] 1 2. [ ] 2 3. [ ] 3 4. [ ] 4
C7.	Is there a control group in the study (receives no treatment at all)	3. [ ] Yes 4. [ ] No
C8.	What is the amount, length, frequency, or dose of treatment	____________________________________ __
C9.	How much time has passed since the intervention occurred?	____________________________________ __
C10.	Does this study focus on equity?	1. [ ] Yes, sub-group analyses are available 2. [ ] Yes, program is targeted 3. [ ] Yes, program is targeted & sub-group analyses are available 4. [ ] No, no equity focus 5. [ ] N/A
C11.	If program is targeted, which sub-group?	1. [ ] low-income students 2. [ ] females 3. [ ] other (describe)

SECTION D. DATA COLLECTION

D1.	What is the sample size of the study?	____
D2.	What is the sample population?	____
D3.	At what level was the data collected?	1. [ ] student/individual level 2. [ ] classroom level 3. [ ] school level 4. [ ] village level 5. [ ] district level 6. [ ] other level (describe ________________________)
D4.	Years over which data was collected	____
D5.	Was IRT used in data collection?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
D6.	Was any data imputed?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
D7.	Was baseline data collected?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
D8.	Was follow-up data collected?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
D9	Was panel data collected?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
D10.	Was cross-sectional data collected?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
D11.	Was repeated cross-sectional data collected?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
D12.	Any comments on data collection/ data type?	____

Section E. Methodology

E1.	Is linear modeling employed?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E2.	Is HLM employed?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E3.	Are quantile regressions employed?	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E4.	Other comments on model	____
E5.	Main identification strategy	____
E6.	Identification strategy is a randomized controlled trial	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E7.	Identification strategy uses instrumental variables	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E8.	Identification strategy uses propensity score matching	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E9.	Identification strategy uses a regression discontinuity methods	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E10.	Identification strategy uses a difference-in-difference methodology	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E11.	Identification strategy uses a panel data/ time series methodology	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E12.	Identification strategy is quasi-random/ a natural experiment	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E13.	Identification strategy uses Heckman selection	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E14.	Identification strategy used does not match any of the above	1. [ ] Yes 2. [ ] No 3. [ ] Not stated/ unclear
E15.	Description of "other" identification strategy *(from E14)*	____

SECTION F. FINDINGS

NOTE: Section F. should be filled out as many times as there are effect sizes estimates

F1.	What is the id number of this impact estimate? (there can be many estimates per study)	____
F2.	What is the name of the student assessment (if this impact is assessment-related)	____
F3.	What is the subject of the assessment linked with this impact?	1. [ ] Math 2. [ ] Language 3. [ ] Science 4. [ ] Composite Score 5. [ ] Other (describe ___________________________)
F4.	What is the education outcome linked with this impact	1. [ ] Achievement 2. [ ] Enrollment 3. [ ] Completion 4. [ ] Retention 5. [ ] Repetition 6. [ ] Drop-Out 7. [ ] Attendance 8. [ ] Teacher Attendance 9. [ ] Other 10. [ ] Other education outcomes (describe:___________________________________) 11. Other non-education outcomes (describe:___________________________________) *If the *only outcomes* covered in the paper are *non-education outcomes* >> EXCLUDE. ◈
F5.	What is the student education level linked with this impact?	1. [ ] Early Childhood Development (pre-K) 2. [ ] Kindergarten 3. [ ] Primary 4. [ ] Middle 5. [ ] Secondary 6. [ ] Primary through Secondary 7. [ ] Tertiary 8. [ ] Other (describe) ____________________________
F6.	What is the student grade level linked with this impact?	1. [ ] ECD 2. [ ] Kindergarten 3. [ ] Grade 1 4. [ ] Grade 2 5. [ ] Grade 3 6. [ ] Grade 4 7. [ ] Grade 5 8. [ ] Grade 6 9. [ ] Grade 7 10. [ ] Grade 8 11. [ ] Grade 9 12. [ ] Grade 10 13. [ ] Grade 11 14. [ ] Grade 12 15. [ ] Grade 13 16. [ ] University 17. [ ] Other (des)

F7.	What is the student sub-group linked with this impact?	1. [ ] All groups/ students 2. [ ] Female 3. [ ] Male 4. [ ] Minority 5. [ ] Minority female 6. [ ] Minority male 7. [ ] Non-minority 8. [ ] Non-minority female 9. [ ] Non-minority male 10. [ ] Low SES 11. [ ] High SES 12. [ ] Low-performers 13. [ ] High-performers 14. [ ] African American 15. [ ] Latino American 16. [ ] ELL 17. [ ] Non-ELL 18. [ ] Special Education 19. [ ] Non-special education 20. [ ] Rural 21. [ ] Urban 22. [ ] Poor health 23. [ ] Other (describe: ________________________)
F8.	What is this impact estimate? (no units)	
F9.	In what units is this impact recorded?	1. [ ] standard deviations 2. [ ] percentage points 3. [ ] percentile points 4. [ ] points 5. [ ] percent proficient 6. [ ] percent 7. [ ] years 8. [ ] days 9. [ ] probit coefficient 10. [ ] logit coefficient 11. [ ] other (describe: ________________________) 12. [ ] not reported
F10.	What is the standard error of this impact?	____
F11.	Describe any calculations used to compute the standard error of the impact (if applicable)	____
F12.	What is the p-value of this impact?	1. [ ] 1% 2. [ ] 5% 3. [ ] 10% 4. [ ] not significant 5. [ ] not reported

F13.	What is the interpretation of the impact? *(i.e. for each additional unit of X, variable Y increases by Z units).*	____ ____ ____ ____
F14.	What is the source of the impact? (regression table or page number)	____ ____
F15.	What estimand is this impact estimate is trying to evaluate?	1. [ ] ITT (Intention to Treat) 2. [ ] ATE (Average Treatment Effect) 3. [ ] ATT (Average Treatment on the Treated) 4. [ ] ATC (Average Treatment on the Control) 5. [ ] LATE (Local Average Treatment Effect) 6. [ ] CACE (Complier Average Causal Effect) 7. [ ] TOT (Treatment on the Treated) 8. [ ] Other (describe: ______________________)
F16.	What is the R-squared associated with this impact estimation?	____
F17.	How many control variables can be found in the regression table –for this specific impact estimate?	____ (describe: ________________________________)
F18.	Any additional comments on this impact estimate?	____
F19.	What is the effect size linked to this impact? (no units)	
F20.	What is the standard error of the effect size linked to this impact?	____
F21.	What calculations were used to estimate this effect size statistic? – as well as the standard error of the effect size.	____ ____