Panel Submission Information

Contact: Kate Schwartz, New York University
kate.schwartz@nyu.edu

Title: Toward contextually-based evidence in international education: Moving beyond what works to questions of for whom and under what conditions

First choice conference section: Educational Effectiveness in Global and Immigration-Related Contexts

Second choice conference section: Research to Practice

Panel Justification:

According to the International Initiative for Impact Evaluation repository there have been 664 impact evaluations conducted in over 120 low- and middle-income (LAMI) countries within the education sector in the past three decades: 61% of which have been since 2010. Much of the data from these trials remains unanalyzed, especially in relation to variation in and mechanisms of program impact. Only between one third and three quarters of impact evaluations across research sectors present the distribution of main program effects (IEG, 2012). Far fewer examine for whom or under what conditions or how programs work best.

As a result, these impact evaluations have not produced clear policy implications. In a meta-analysis of interventions targeting learning outcomes in developing countries, McEwan (2015) found that for nearly half (5/11) of the intervention types (e.g., teacher training), there was more variation between programs in the same category than across programs in all categories. This variability within ‘like’ interventions, combined with a lack of understanding of what produces it, suggests that researchers are not capitalizing on the large investment in these trials toward policy- and practice-relevant knowledge. Evaluating an intervention costs $500,000 on average, approximately 1.4% the price of the intervention (IEG, 2012; Piccio, 2014). This equates to over 200 million dollars spent on education impact evaluations (and around 14.2 billion on accompanying interventions) in LAMI countries in the past six years. Given the large variation found in evaluations to date and the high amount of need in LAMI countries, much more needs to be done with these investments in order to garner information that can effectively be used to target services, programs, and policies within the education sector.

There are practical barriers to fulfilling this prescription. In this panel, we first present an overview of how often questions of variation in and mechanisms of impact are asked within education research in the LAMI context. This presentation will include a discussion of potential reasons such questions are not asked more frequently and ways to move the field in the direction of more nuanced, policy-relevant inquiry. We then present two papers that delve into the challenges related to statistical power and measurement of outcomes. The first of these examines the role of research design and statistical power on the ability to answer questions of variation in program impact. The second presents an example of work seeking to build a measure of early
childhood development that is valid within diverse, local contexts and comparable across contexts. Such work, especially when applied to settings-level measures where smaller sample sizes make power a much more larger concern, could allow for explorations of for whom and under what conditions programs work even when individual studies are under-powered. In addition, establishing statistically invariant measures and using them across studies allow for valid comparisons of program effectiveness. We conclude with a discussion led by Dr. Michael Weiss, an expert in variation in impact within the U.S. context and author of a 2014 theoretical paper on sources of variation in program effects.

Paper #1

Title: Taking Stock: Frequency of, Barriers to, and Potential Ways to Foster more Policy-Relevant Education Research in Low- and Middle-Income Countries

Presenting Author: Kate Schwartz, New York University kate.schwartz@nyu.edu

Authors: Kate Schwartz, New York University; Yeshim Iqbal, New York University; J. Lawrence Aber, New York University

Background: There has been an exponential growth in impact evaluations in LAMI countries. In the education sector, nearly two thirds of all published impact evaluations, according to 3ie’s online repository, have been conducted in the past 6 six years. Yet it is unclear what we are learning about effective education policies and interventions from such work. Meta-analyses and systematic reviews of education research in developing countries have found huge ranges in effectiveness even within conceptually similar interventions (Conn, 2014; McEwan, 2015). Despite this, very few impact evaluations – across all research sectors – examine questions related to variation and mechanisms of program impact (IEG, 2012). Answering these kinds of questions is a critical component of effectively scaling and replicating programs and central to any efforts to match local needs to possible education interventions. Given this, and the large amount of resources currently being invested in educational impact evaluations in LAMI countries, it is increasingly important that we maximize knowledge gained from work already conducted and orient future research so as to facilitate answering nuanced questions about for whom and under what conditions programs work. In order to do so, we must first understand exactly how frequently such questions are or are not answered within the education sector and what barriers and supports contribute to the frequency or infrequency of such lines of inquiry.

Purpose/Objective/Research Question: This paper provides a ‘state of the field’ for educational intervention impact evaluations in LAMI countries. Utilizing quantitative analyses of published papers, working papers, and online reports, we descriptively examine what kinds of research has been conducted, where, and with what populations within the education sector in LAMI countries with a focus on the types of research questions asked and answered. We then
discuss possible reasons why questions pertaining to variation and mechanisms of program impact are not addressed more frequently. For this discussion, we draw on qualitative data collected through surveying researchers working in LAMI contexts. The goal is to provide a foundation for better understanding current education research in LAMI countries and for aiding the field as it seeks to advance into more frequent, nuanced inquiry regarding how, for whom, and under what conditions programs work.

**Research Design:** We systematically searched the World Bank, Innovation for Poverty Action, and 3ie data repositories to find relevant evaluations. We then assessed each evaluation to ascertain the extent to which each one: i) examines the distribution of program impacts; ii) explores differential impacts by individual characteristics such as gender or ethnicity; iii) moves beyond these demographics to examine other individual moderators; iv) assesses impacts by local setting characteristics such as rural versus urban or resource availability; and, v) tests mechanisms through which the program worked.

For the qualitative component of our analyses, we created a survey to assess researchers’ perceptions of both their own and their field’s frequency of asking and answering questions around variation and mechanisms of program impact as well as their thoughts as to what some of the barriers to such work might be. This survey was distributed via Innovations for Poverty Action, 3ie, the World Bank, and Global TIES for Children research affiliates. We have had 21 researchers respond to data and are currently both still actively soliciting and receiving responses. Responses will be analyzed using thematic analysis of open-ended questions as well as frequency counts for a handful of multiple choice items.

**Findings/Results:** In analyzing published impact evaluations, working papers, and online reports in the education sector, we find that the plurality deal with providing resources for low-income students (e.g. conditional cash transfers and school feeding programs). The majority of research to date has taken place in Mexico, India, Colombia, Kenya, China, Chile, and Indonesia. Around 50% of impact evaluations in education are randomized control trials, with the rest being one of or some combination of difference-in-difference, propensity score, instrumental variable, and regression discontinuity designs. While many programs (42.0%) examine differential impacts by individual characteristics (for whom?) very few (19.7%) do so by settings-level characteristics (under what conditions?) and ever fewer (13.5%) look at mechanisms through which the program worked. Furthermore, the majority of evaluations that examine individual moderation do so only by one (and in some cases two) of the following variables: gender, baseline performance of the key outcome variable, or age of student. For those studies that do examine settings level moderation, around half do so by geographic region, not any specific characteristic of the setting itself.

**Conclusions:** The above results will be discussed in greater detail along with our findings from the researcher survey currently being collected. We will discuss how the infrequency of inquiry effects research replication, program scalability, policy, and practice across contexts and populations. We will also present our hypotheses for why inquiry surrounding for whom, under what conditions, and how program work is so infrequently seen in research broadly and education research specifically. This discussion will include potential barriers to both knowing if such questions were posed (i.e. publication bias) and asking/answering such questions (i.e. power, measurement). We
close with a discussion of how, as researchers focused on improving policy, we might better leverage existing data and better design future evaluations in order to generate more policy-relevant, actionable, context-informed evidence. This discussion, as to both the barriers to such inquiry and potential solutions, will be extended in the two papers that follow.


**Paper #2**

**Title:** An Examination of the Design and Statistical Power of Impact Evaluations in Low- and Middle-Income Countries

**Presenting Author:** Jessaca Spybrook, Western Michigan University

**jessaca.spybrook@wmich.edu**

**Author:** Jessaca Spybrook, Western Michigan University; Qi Zhang, Western Michigan University; Dustin Anderson, Western Michigan University

**Background:** The frequency of rigorous impact evaluations in low- and middle-income (LAMI) countries is on the rise. Since 2010, nearly 400 impact evaluations related to children’s educational outcomes have been conducted in LAMI countries (Schwartz, Iqbal, & Aber, paper 1 of panel). Many of these impact evaluations rely on randomized trials, and specifically cluster randomized trials (CRTs). For example, in the impact studies focused on educational interventions, the focus of this paper, interventions are often implemented at the school or village level and hence the natural unit of random assignment is the cluster. Students are nested within schools or villages yielding a 2-level CRT.

Most of these evaluations are designed with the *what works* question in mind. Hence the evaluation is powered to detect a main effect of treatment. However, program effects may differ by individual characteristics, group characteristics, or setting-level features. Given this, questions related to *for whom* and *under what conditions* may also be of importance. If this is the case, then it becomes important to consider these types of moderator effects in the planning phase of the study.

**Purpose/Objective/Research Question:** The purpose of this paper is to examine the capacity of impact evaluations of educational interventions in LAMI countries to provide rigorous evidence to important policy questions related to *what works*, *for whom*, and *under what conditions*. Using a
subsample of the studies identified by Schwartz, Iqbal, & Aber (described in paper 1), we explore
the capacity of the studies to answer 1) the *what works* question by estimating the minimum
detectable effect size (MDES), 2) the *for whom* question by estimating the minimum detectable
effect size difference (MDESD) for an individual level moderator, and 3) the *under what
conditions* question by estimating the MDESD for a cluster level moderator. The findings from this
study will help us better understand the potential of these studies to generate rigorous evidence
related to these three questions. Further, the data from these large impact studies is often made
publically available and knowing the precision with which these studies are designed will help the
field better understand the potential for conducting meaningful secondary data analyses.

**Research Design:** From the set of studies identified by Schwartz, Iqbal, & Aber, we identified
those studies that meet the following criteria: conducted after 2009, focused on evaluating an
educational intervention, utilize a CRT in which schools or villages are the unit of
randomization, and outcome data is available at the student level. Next we confirmed the specific
type of CRT and the unit of random assignment. For example, 2-level CRTs include studies with
students nested in schools and schools randomly assigned to condition, 3-level CRTs include
studies with students nested classrooms nested in schools and schools randomly assigned to
condition, and multisite CRTs include studies with students nested in classrooms nested in
schools but classrooms are randomly assigned to condition within schools. Then we will
determine the sample sizes at each level and empirical estimates of appropriate design
parameters. Based on this complete information, we plan to estimate the 1) MDES (formula
provide Bloom, 2005), 2) MDESD for individual level moderator effects (formulas provided by
Spybrook, Kelcey, & Dong, forthcoming), and 3) MDESD for cluster level moderator effects
(formulas provided by Spybrook, Kelcey, & Dong, forthcoming). We will use PowerUP!
(www.causalevaluation.org) to operationalize the formulas.

**Findings/Results:** The full results will be provided in the study. The results will include the
number of studies for each design category, the number of clusters randomized separated by type
of cluster (school or village), and a range for the MDES and MDESD for the individual and
cluster level moderator for each study. A range will be presented as we will be using a range for
empirical design parameters to estimate the lower and upper bound for each parameter.

We provide a sample of the results for 4 studies (see Table 1).

<table>
<thead>
<tr>
<th>Design Classification</th>
<th>Cluster Randomized</th>
<th>Total Number of Clusters Randomized</th>
<th>MDES</th>
<th>MDESD Individual level moderator</th>
<th>MDESD Cluster level moderator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>2-level CRT School</td>
<td>100</td>
<td>0.22</td>
<td>0.22</td>
<td>0.435</td>
</tr>
<tr>
<td>Study 2</td>
<td>2-level CRT School</td>
<td>200</td>
<td>0.15</td>
<td>0.16</td>
<td>0.306</td>
</tr>
<tr>
<td>Study 3</td>
<td>2-level CRT School</td>
<td>93</td>
<td>0.23</td>
<td>0.23</td>
<td>0.45</td>
</tr>
<tr>
<td>Study 4</td>
<td>2-level CRT School</td>
<td>85</td>
<td>0.24</td>
<td>0.24</td>
<td>0.47</td>
</tr>
</tbody>
</table>

*Calculations were based on the following assumptions: two tailed test, alpha = 0.05, 20 students per school, cluster
level covariate that explains 30 percent of the variation, intraclass correlation of 0.15, equal allocation at all levels, individual level covariate that explains 30 percent of the variation.

**Conclusions:** How much can we learn from the rigorous impact evaluations of educational interventions in LAMI countries? How much can we learn from secondary analyses of the data from these large intervention studies? One way to assess this is to examine the capacity of the studies to answer questions related to what works (or the main effect of treatment), for whom (or individual moderator effects), and under what conditions (or cluster moderator effects). Findings from this study will shed light on the precision associated these different effects.


---

**Paper #3**

**Title:** Establishing measurement invariance of the International Development and Early Learning Assessment (IDELA) across five countries

**Presenting Author:** Peter Halpin, New York University

peter.halpin@nyu.edu

**Authors:** Sharon, Wolf, University of Pennsylvania; Peter Halpin, New York University; Hiro Yoshikawa, New York University; Natalia Rojas, New York University; Sarah Kabay, New York University; Amy Jo Dowd, Save the Children; and Lauren Pisani, Save the Children

There is increasing interest by researchers and policy makers around the world to improve children’s early skills and knowledge. With the release of the Sustainable Development Goals in 2015 (UN, 2015), Target 4.2 under Education Goal 4 aims to ensure equitable learning for all, and indicator 4.2.1 specifically calls for the development and utilization of cross-country comparable measures that encompass multiple domains of early childhood development (ECD). In addition, researchers are often interested in comparing impacts of programs across different contexts.
The research base and policy context of current efforts in international education has brought to the forefront the importance of developing assessments of ECD and school readiness that are feasible to administer, conceptually and psychometrically validated across contexts, aligned with national monitoring systems, and are comparable across countries and contexts. The IDELA is a direct child assessment developed by Save the Children to assess early childhood development holistically across low- and middle-income countries (LMICs) (Pisani et al., 2015). The tool has been used in over 30 countries and is drawing growing attention from the research and donor communities. This study is the first to examine the psychometric properties of the IDELA across five LMICs (Afghanistan, Bolivia, Ethiopia, Uganda, and Vietnam).

**Measures:** All children were assessed using the IDELA assessment, which was translated into the relevant local languages of the mother tongue of the participants, adapted using a process of review and field testing by each country’s Field Office in collaboration with the tool developers from Save the Children, U.S. For example, in developing the letter identification subtask, high and medium frequency letters were chosen in country with local educational staff. The assessment takes an average of 35 minutes/child to administer.

The tool measures four domains of development. *Motor development* consists of 10 items, grouped into 4 subtasks, assessing both gross and fine motor skills. *Social-emotional development* consists of 14 items grouped into 5 subtasks, and covers constructs of self-awareness, emotion identification, perspective taking and empathy, friendship, and conflict/problem solving. *Early literacy* consists of 38 items grouped into 6 subtasks, and covers constructs of print awareness, letter knowledge, phonological awareness, oral comprehension, emergent writing, and expressive vocabulary. Finally, *early numeracy* consists of 39 items grouped into 8 subtasks and covers constructs of number knowledge, basic addition and subtraction, one to one correspondence, shape identification, sorting abilities based on color and shape, size and length differentiation, and completion of a simple puzzle.

**Analytic Plan and Preliminary Results:** Based on the theorized four-domain structure of the IDELA, as well as the presence of subtasks within each domain, we hypothesized that the IDELA could be modeled using a bi-factor approach within each domain. In order to test this hypothesis, each country dataset was randomly split into two subsamples, which we refer to as the exploratory samples and the confirmatory samples (see Table 1 for power analysis used to determine sufficient sample size).

All of the analyses were conducted in two steps. First we used the exploratory samples to establish a plausible bi-factor structure within each domain, within each country. We also assessed various models for the overall structure of IDELA instrument, in which the relations among the four factors were examined. Next, the resultant within-domain models and the overall models for the IDELA were tested in the confirmatory samples within each country. This facilitates model development and provides initial evidence on the out-of-sample generalizability of the proposed models. Results support the conclusion that the four domains can be adequately represented using a single common factor, with a bi-factor/testlet structure to model local dependence of items on the same subtask. Table 2 presents within-country goodness of fit statistics for the domain-specific confirmatory bi-factor models.
In our next steps, we will use measurement invariance analyses to assess the degree to which cross-country comparisons on the four domains are warranted. The within-country analyses provide a basis for expecting the configural invariance model to hold (i.e., same number of factors, with the same items loading on the same factors, in all five countries). However, it would be optimistic to expect that all of the items are invariant over all five countries. Because we anticipate a large degree of measurement non-invariance (i.e., item bias) over countries, our approach to assessing measurement invariance deviates from the standard method of testing progressively more restrictive parameter constraints on the configural model (e.g., see Millsap, 2012). Rather, we will start our cross-country analysis of the exploratory samples with the most restrictive model (scalar invariance, which is required for group mean comparisons over countries to be warranted), and examine the univariate and bivariate residuals of each item from this model, within each country. On the basis of these exploratory analyses, we will develop an overall structural model describing the invariant and non-invariant components of the IDELA, and test this model in the confirmatory samples. Our final results will be presented in terms of which of the items on each domain of IDELA are useful for making comparisons over countries.

**Conclusion:** In the post 2015 context for international development, assessments of early childhood development are increasingly in demand in LMICs. While several tools are currently being developed and used, most of these have not been subject to even preliminary psychometric analyses. This is the first study assessing the psychometric properties of a direct early childhood assessment across multiple LMICs. The analyses provide insight into the promise of using factor analytic methods to understand the utility of measures developed to assess children’s school readiness, and lay the groundwork for similar research with the other types of measures. The findings will also provide empirical evidence of the utility of the IDELA for use within each country (e.g., for program evaluation, for reporting subgroup differences such as gender and region), and for its potential for use in evaluating and comparing impacts of ECD programs across countries. This study provides a critical step in understanding how results from global assessment tools and results from different policy efforts can be compared across countries.


