Methodology for SkillsBank (version 2017-08-25)

Introduction

This document describes the methodology used to select the evaluations, code the variables and generate the results presented in the SkillsBank. Four different reviews were performed and Table 1 lists them, along with the outcomes that they covered. Notice that certain reviews analyzed interventions focusing on one outcome while other reviews analyzed interventions affecting different outcomes. For example, the Learning in Primary review analyzed evaluations of interventions focusing on learning outcomes. But, the Early Childhood review analyzed interventions that sought to improve early childhood cognition or early childhood behavior.

Table 1: Reviews and outcomes

Review	Outcomes
1. Learning in Primary (**benchmark**)	Learning in Math and Language
2. Early Childhood	Early childhood cognition Early childhood behavior
3. Secondary Enrolment and Completion	Enrolment in secondary Completing secondary
4. Learning in Secondary	Learning in Math and Language

The reviews followed similar but not identical methodologies. There are advantages of following similar methodologies related to ensure consistency of the results and adherence to benchmark procedures. But, there are certain methodological aspects that will vary across reviews because they are directly linked to the nature of the interventions. For example, the categories of interventions in the Learning in Primary review and in the Early Childhood review vary, because the interventions in both reviews are quite different. Moreover, there are also advantages of allowing departures from the benchmark methodology in other aspects that could be kept fixed across reviews (e.g., the search strategy and the inclusion criteria imposed) to ensure that each review can be adapted to the specificities prevalent in each case. We balance these objectives by seeking to follow a benchmark methodology but allowing departures from it when we considered that it was convenient to adapt the methodology to specific situations.

The departures from the standard methodology are more common during the initial part of the process of each review. This is because the search strategies and the inclusion criteria of the studies draw on the availability of the existing reviews and studies in the literature. For example, for the Learning in Primary review we did not consider that it was necessary to do a keyword search in bibliographic databases, such as google scholar, because there were several recent reviews that could be used to identify potentially relevant studies. In contrast, for the review regarding early childhood development, it was decided that a systematic search in bibliographic databases was necessary because there were not many updated recent reviews available. However, in the final steps of the reviews, involving aggregation of the effect sizes, we considered that the gains of standardization outweigh the potential gains of following different strategies. Consequently, all reviews follow the same methodology during this stage of the process.

In the remainder of this document, we first describe the Learning in Primary review because this review was used as a benchmark. We then describe how the other reviews deviate in certain methodological aspects such as the inclusion criteria from this benchmark.

1. The benchmark: Learning in Primary review

A systematic review involves performing a number of steps that are summarized in Table 2. These steps are described in more details in the next subsections.

Table 2: Steps followed in the review

Steps	Description
1. Search papers	Identify papers reporting evaluations that can potentially be included in the review
2. Filter evaluations	Keep only evaluations that fulfill the inclusion criteria
3. Code variables	Extract and code relevant variables from the included evaluations
4. Categorize interventions	Classify categories in groups to define program types (e.g., lesson plans)
5. Compute effect sizes	Compute the effect sizes
6. Combine effects sizes	Combine effects to produce one summary effect per evaluation
7. Generate averages	Analyze data and produce overall effect by running a meta-analysis

1.1. Search papers

The first step in the review involved searching for studies that could be potentially included in the review. To do this search, there are two main approaches that can be used. The first approach involves identifying potential studies among those cited in the recent reviews in the literature. The advantage of this approach is that, because it builds on the work done in previous reviews, the number of potential studies to review is reduced. A second approach involves performing keyword searches in bibliographic databases. This approach can also yield potential studies to be reviewed, but it is highly intensive in terms of time because it typically requires reviewing thousands of studies. Moreover, many studies may not be identified using this approach because they do not use the words included in the strings searched. Additionally, there are other approaches such as contacting experts in the field to suggest studies, reviewing specific journals or working paper series that tend to publish studies in the area. Finally, potential studies can be identified by checking the studies that cite or are cited in those found in the first stage.

For the Learning in Primary review we noticed that there were several recent, well-executed reviews in different strands of the literature that could be used as main sources for identifying studies. Table 3 presents the existing reviews that were used as inputs for this process for developed countries and Table 4 presents the reviews for developing countries. Consequently, our main approach was to identify studies that were included in these reviews.

Table 3: Existing reviews – Developed countries

Authors	Only RCTs?	Studies
Fryer (2016)	Y	196
Cheung et al. (2013)	N	74
Rakes et al. (2010)	N	82
Slavin et al. (2009)	N	62
Slavin et al. (2007)	N	87

Table 4: Existing reviews – Developing countries

Authors	Only RCTs?	Studies
McEwan (2015)	Y	77
Glewwe and Muralidharan (2016)	N	115
Murnane and Ganimian (2014)	N	115
Glewwe et al. (2011)	N	79
Kremer et al. (2013)	Y	30
Krishnaratne et al. (2013)	N	75

The references of these studies were located and duplicates were dropped. The resulting database included all studies that could potentially be included in the review.

1.2. Filter evaluations

The second step in the review involved determining which of the studies identified in the search step should be included in the review. To that end, it was necessary to define inclusion criteria with the requirements that a study had to fulfill to be included in the review. Defining these requirements typically involves balancing two competing objectives: reducing bias and increasing precision. For example, if we only include experimental studies, we likely would be reducing biases due to the recognized ability of this design methods of minimizing selection issues. However, it could be the case that for certain interventions there would be only one or two randomized evaluations (or not even one, such as in the case of extending the school day). Consequently, imposing this stringent methodological requirement would likely reduce the precision of the effects estimated in the review.

In general, because the literature on how to improve learning in primary education is mature, we could adopt quite stringent requirements for studies to be included. For making these decisions, we analyzed the requirements chosen in prior studies. In particular, a major input for our definition of the inclusion criteria for this review were the studies produced by Slavin and co-authors that dealt with a number of important issues, including issues directly related to the methodological design used to deal with selection issues, but also other aspects that are less recognized though they seem to be very important (e.g., ensuring that the tests used are "balanced" in the sense that they include material that was covered both in the treatment and control groups).

The requirements included in the inclusion criteria are reported in Table 5.

Table 5: Inclusion criteria

Requirements
1. The evaluation was implemented in primary or elementary school
2. The intervention aimed to improve learning in math or reading
3. The evaluation compared the treatment group to the status quo
4. The effects were measured at least 3 months after the start of the intervention
5. The effects were estimated using RCT, RD, IV or differences-in-differences
6. The tests measured content instructed in both the treatment and comparison groups
7. The effects were measured using a continuous measure of learning
8. The sample included at least 200 students and 10 groups (e.g., schools)*
9. The standard errors were computed incorporating clusters when necessary
10. Sufficient information was reported to compute effect sizes

Notes:
* The group requirement was applied if randomization was performed at the group level.

Moreover, evaluations tend to report multiple effects and we followed these criteria to determine the effects to record:

(i) The effect for the first follow-up performed after 12 weeks of intervention was extracted.

(ii) Effects for a summary academic achievement in Math and Language (usually the effect on the average score) were prioritized. If not available, then effects on Math and Language were extracted.

(iii) Effects for only one specification were extracted following pre-established criteria.

(iv) If an evaluation reported effects from different research designs, we extracted the one preferred by the author and if there were no one that was suggested as preferred, then we extracted all of them.

1.3. Code variables

The third step in the review involved coding variables relevant for the analysis. The variables were extracted at two levels: evaluation and effect. That is, there are certain variables that are defined at the evaluation level. For example, the country where the intervention was implemented is defined at the evaluation level. In contrast, other variables are defined at the effect level. That is, an evaluation can report multiple effects (e.g., one for Math and another for Reading) and hence certain variables are defined at this level. Table 5 describes the variables extracted and coded at the evaluation level. Table 7 describes the variables extracted and coded at the effect level.

Table 6: Variables coded for each evaluation

Variable	Description
Evaluation	Last name of first author of the paper reporting results followed by the year of the publication*
Effect	Summary effect at the evaluation level (obtained from a meta-regression)
Average age	Average age at baseline; if only grade information reported, then, average grade
Observations	Number of observations of the effects included
Experimental Evaluation	Indicator for experimental evaluations
Country	Country were the evaluation was implemented
Intervention	Intervention description (short)
True effect lies - low value	Lower limit of 90% confidence interval
True effect lies - high value	Higher limit of 90% confidence interval
Weight in average effect (%)	Weight of the evaluation to compute aggregate effect (obtained from a meta-regression)
GDP per capita	GDP per capita in the year when the intervention evaluated started (in 2015 dollars)
Age	Age of participants in the evaluation
Context details	Description of the context of the evaluation
Implementation year	Year when the intervention evaluated started
Duration (months)	Months elapsed between the start of the intervention and measurement of effects**
Delay (months)	Months between the end of the intervention and measurement***
Implementer	Government, NGO, Private, Researchers or Other
Cost	Information on costs reported in the evaluation
Intervention details	Intervention description (detailed)
Evaluation design	Randomized evaluation, Regression discontinuity, Instrumental variables, Differences-in-differences
Cluster definition	Individual, Class, School, Geographic area, Center, Other
References	Reference of the main study formatted using Chicago Manual of Style

Notes:
* If one paper reports multiple evaluations, then roman numerals are added. For example, the evaluations Worth 2015 i and Worth 2015 ii are reported in the paper Worth et al. (2015).
** If measurement took place after the end of the intervention, then duration corresponds to the number of months between the start and the end of the intervention.
*** If measurement took place before the end of the intervention, then it takes a value of 0.

Table 7: Variables coded for each effect

Variable	Description
Sample	Indicates the sub-sample for which the effects correspond (e.g., "Grade 3")*
Outcome	Outcome measured (e.g., Math)
Test	Test used to measure effects
Source table	Table in the paper that reports the effect extracted
Observations	Number of observations included in the effect extracted
Statistics on effects	Statistics used to compute effect sizes

Notes:
* If all individuals in the main sample were included, then it indicates "Complete sample."

1.4. Categorize interventions

The fourth step in the review involve categorizing interventions in groups. We call these groups “program types.” The basic goal of this step is to group similar interventions together. In our analysis, we generate categories that are exhaustive and mutually exclusive. That is, all interventions should be assigned to one, and only one, program type. Moreover, we define quite narrower categories to increase the similarity between interventions assigned to the same program type. Specifically, we define 20 program types that include, for example, categories of interventions such as “monetary incentives to students” or “lesson plans.” Table 8 presents 20 program types in which the interventions in the Primary in Learning review were categorized.

Table 8: Program types

Program types	Description
Books	Books or related services (e.g., libraries or librarians)
Class size	Reduction of the number of students in a classroom from 25 to 20
Community involvement	Training and budget to promote community involvement
Computers	Computers and other devices such as laptops and tablets
Funding for materials	Small budget to principals or teachers
Guided technology (no extra time)	Instruction using technology during regular time
Guided technology (with extra time)	Instruction using technology after hours
Lesson plans	Lesson plans to teachers to facilitate instruction
Managers development	Training or coaching to principals or district managers
Monetary incentives to students	Monetary incentives to students based on performance
Monetary incentives to teachers	Monetary incentives to teachers based on performance
Non-monetary incentives to students	Non-monetary incentives to students to promote effort
Parent training	Training to parents through group meetings or home visits
School day length	Extending the school day from 4 to 7 hours
Small group support	Instruction to students in small groups
Teacher development	Training or coaching to teachers to promote better instructional practices
Teachers' years of education	Increasing teachers' years of education by 2 years
Test scores data	Information about students' academic achievement to principals or teachers
Tracking	Assignment of students to instructional groups based on achievement
Tutoring	One-to-one 30-60 min instruction of students 1-5 times a week

1.5. Compute effect sizes

The fourth step in the review involved the computation of effect sizes. An effect size is a quantitative measure of the effect of an intervention on certain outcome that is comparable across evaluations. Because the effects are comparable, they could be used to estimate an overall summary effect.

For the Learning in Primary review, we use as effect sizes the standardized mean difference (*d*). The standardized mean difference can be computed by dividing the difference between the mean of the outcome in the treatment and the control group by the (within) standard deviation of the outcome. That is:

$$d = \frac{Mean_{treatment} - Mean_{control}}{SD_{control}}$$

Moreover, it is an increasingly prevalent practice in this literature to report effects in terms of the standardized mean difference. Consequently, the reported results are already expressed in terms of this effect size. To present effects in terms of the standardized mean difference, primary studies transform the outcome from raw scores to standardized scores using the following formula:

$$Y_{st} = \frac{Y_{raw} - Y_{mean}}{Y_{sd}}$$

where Y_{st} is the standardized score, Y_{raw} is the original (raw) score, Y_{mean} is the mean value for the raw score and Y_{sd} is the standard deviation of the raw score.

Still, there are certain studies that do not present effects in terms of standardized mean differences. In these cases, we use other statistics to compute the standardized effect size. For example, for a study that reports post treatment means of not standardized tests scores for treatment and control groups and standard deviation for the control group test score after treatment, we compute the standardized effects sizes as a difference between post treatment means divided by the control group standard deviation.

Finally, for two program types – class size and school day length – we standardize the effects so that all evaluations present effects of a similar "intensity" of the intervention. First, there are several evaluations of class size which effects correspond to different reductions in class size (e.g., one study may evaluate reduction in class size from 22 to 15, but another one from 40 to 20). We standardize the effects from these evaluations so that all effects correspond to the same reduction in class size from 25 to 20. To do that, we assume that the effects on learning are linear in terms of the percent reduction in class size. Second, for the interventions that extend the school day we standardize the effects so that all effects correspond to the same increase in school day length from 4 to 7 hours. For this standardization, we assume that the learning effects are linear in terms of the increase produced by the intervention in the number of school hours.

1.6. Combine effect sizes

Evaluations routinely report multiple effects regarding how an intervention affect student learning. For example, an evaluation can report multiple effects corresponding to different outcomes such as Math, Reading, or an average of both subjects. Moreover, an evaluation can report multiple effects for one outcome because different tests were applied. Including multiple effects for one evaluation in a meta-regression is problematic because it requires to assume that the included effects are independent. This is clearly not the case if the different effects were obtained from the same sample of participants. There are different approaches to tackle this issue. In our analysis we follow the recommendation of Borenstein et al. (2009) who suggest combining different effects for one evaluation to produce a summary effect.

To combine different effects for one evaluation it is important to recognize whether the effects are expected to be correlated. This is the case, for example, when Math and Reading effects are estimated. Because the same sample of students are tested, these effects are likely to be correlated. Also, the effects are likely to be correlated when they correspond to different tests applied to the same sample of students. Similarly, the effects are likely to be correlated when they correspond to different methodological designs (e.g., experimental evaluations and differences-in-differences) that were applied to measure the effects of the same intervention in the same sample of students.

For these cases, we follow the procedures suggested by Borenstein et al. (2009) for computing combined effects and their standard errors. In particular, we compute the combined effects as means of the original effects sizes and we assume correlation of 0.5 for computation of standard errors (results are robust to choosing different plausible values for this correlation).

A different issue is raised when an evaluation presents results for different subsamples of individuals (e.g., from different grades) but it does not present the effect for the overall sample. That is, if the evaluation reports the results for the overall sample, we extract that effect. But if only results for different sub-samples are presented, then we combine them. Because the effects are estimated from different samples of individuals, we can think that they are independent. In this case, we follow the suggestion of Borenstein et al. (2009) to generate a combine effect by performing fixed-effects meta-regression on the reported effects.

The described procedures produces a unique effect size for each evaluation. Still, there is a final issue that needs to be dealt with. If two evaluations share the same control group and correspond to the same program type (e.g., lesson plans), they will be aggregated in the same meta-regression. This again creates problems because the two effect sizes are not independent. This is because they share the same control group and hence the effects are clearly related (if the average outcome for the control is lower than expected due to sampling variability then the effect will be positive for both evaluations). For this case, Borenstein et al. (2009) suggests combining both effects size into a summary effect that corresponds to a study level (rather than the evaluation level).

For example, the paper by Muralidharan et al. (2011) presents two evaluations: (i) individual incentives for teachers and (ii) group incentives for teachers. The first evaluation is labeled in the SkillsBank as Muralidharan 2011 i and the second evaluation is labeled as Muralidharan 2011 ii. Each treatment in this evaluation is compared against the same status-quo control group. Because both evaluations share a control group and belong to the same intervention, we need to combine them. Thus, we define a study labeled Muralidharan 2011 and compute the effect of this study as a simple average of the effects from each evaluation in this study. We compute standard error assuming correlation 0.5 as suggested by Borenstein et al. (2009).

1.7. Generate averages

The final step in the Learning in Primary review involved generating the overall average effect for each program type. Specifically, we run a random-effects meta-regression for each program type including as the unit of observation each study that measured the effects of all interventions in that type. For example, to compute the overall average effect of the program type “guided technology (with extra time)” we run a random-effects meta regression including the following studies that measured the effects of this type of program: Banerjee 2007 ii, Lai 2013 i, Lai 2013 ii, Lai 2015, Linden 2008 ii and Mo 2013.

Running this random-effects meta regression basically involves computing a weighted average effect of the included studies. The weights for study *i* are computed using the following formula:

$$W_i = \frac{1}{se_i^2 + T^2}$$

where se_i^2 is the standard error of the estimated effect from study *i* (i.e., the variance of the estimate) and T^2 is the between-study variance. Note that standard error of the estimate of each study is a function of the sampling variability in the study. That is, studies with larger sample sizes should have lower standard errors (other things equal). In contrast the between-study variance measures how much the true effects vary across the different studies included. We estimate the between-study variance using a method of moments approach.

The formula for computing weights of individual studies shows that those that are more precise (i.e., with lower standard errors) will have higher weights in the computation of the overall average effect. However, in cases where the estimated between-study variance is larger, the difference in weights for the more precise studies will be attenuated. Still, these studies will have higher weight than studies with larger standard errors (see Borenstein, 2009, chapter 12 for a detailed explanation).

Finally notice that the meta regression is run at the “study level” and not at the “evaluation level.” This is because, as it is noted at the end of the previous section, there are some evaluations whose effects are correlated (i.e., those sharing the same control group and belonging to the same program type). Consequently, these effects are aggregated at the study level and these are the effects that are included in running the meta regression for each program type. Notice that for evaluations that are independent with all the other evaluations in a program type (the majority of them), the evaluation coincides with the study. That is, the effects of the evaluation Duflo 2012 i (that is independent of all other evaluations of teacher incentives) are the same as of the study Duflo 2012 i.

Still, we considered it more informative to presents results in the SkillsBank for individual evaluations rather than studies. That is, returning to the example presented in the previous section, information is presented in the SkillsBank for the evaluations Muralidharan 2011 i and Muralidharan 2011 ii. That is, we present the information separately for both evaluations because we can see that the first one involved individual monetary incentives whereas the second intervention involved group monetary incentives. A final issue is that we are presenting weights for individual evaluations in the graph of effects presented in the SkillsBank. For evaluations that are independent of other evaluations in the same program type (e.g., Duflo 2012 i), the weight of the evaluation is just the weight of the study with the same name. However, for evaluations that are correlated with others in the same program type, we assign the weight for the aggregate study proportionally to the weight of the evaluations in the random-effects meta regression run at the evaluation level.

In summary, the procedure of computing weights for evaluations was the following: (i) perform a random effects meta-analysis including evaluation-level effects for all evaluations in the program type analyzed, (ii) perform a random effects meta-analysis including study-level effects for all studies in the program type analyzed, (iii) compute the evaluation-level weight as a proportion of the study-level weight obtained in step (ii), where the proportions are given by the weights in the meta-analysis performed in step (i).

1.8. Special issues

1.8.1. Terminology

To facilitate the discussion, we define the following terms:

- *Effect size*: a quantitative measure of the effect of an intervention on an outcome. The specific measure is selected so that it is comparable across studies.

- *Evaluation*: an empirical analysis performed to estimate the effect of certain intervention on certain outcome. To estimate this effect, any evaluation will always entail the comparison of a group of individuals that benefitted from this intervention (treatment group) and a group of individuals that did not (control group). Notice that some studies involve multiple treatments. For example, one study may randomize students to a treatment group involving tutoring, another treatment group involving playing academic games in a computer and a control group. For our systematic review we consider that this study includes two evaluations. One involves the comparison of the first treatment group (tutoring) with the control group and the other involves the comparison of the second treatment group (playing academic games in a computer) with a control group. Notice that we do not include as an evaluation the comparison of the two treatment groups.

- *Intervention*: a purposeful change of the status quo introduced to improve certain outcome. Typically, it involves a range of activities. For example, interventions that involve providing computers to schools also typically include the provision of software installed in the computers, teacher training and technical support.

- *Paper*: any document that contains information about an evaluation. It includes a variety of documents such as papers published in peer-reviewed journals, working papers, chapters in books, mimeographs and government reports.

- *Program type*: a category of interventions that are similar. For example, in the Learning in Primary review we define monetary incentives to teachers as a program type.

- *Study*: an empirical analysis performed to estimate the effect of certain intervention on certain outcome. In cases in which an evaluation does not share the control group with other evaluations in certain program type, the evaluation coincides with the study. However, if there are two evaluations that share a control group and are classified in the same program type, then the effect sizes of both evaluations are aggregated to create only one effect size at the study level.

1.8.2. Multiple reports of the same evaluation

In many cases, researchers produce a working paper of an evaluation and later on the evaluation results are reported again in a journal article. In these cases, we prioritize extracting results from the most recent published version of the evaluation (typically the journal article) because we can expect that this version should incorporate a number of potential improvements identified since the publication of the first report and the second report.

2. Early Childhood review

This section presents the departures from the benchmark methodology that occurred when producing the Early Childhood review.

2.1. Search papers

For this review, we searched studies evaluating parenting programs worldwide. In particular, we searched peer reviewed articles in the following digital databases: Cinhal, Medline, PsycINFO and ScienceDirect. We also included relevant citations of those papers and papers from other sources including expert recommendations.

2.2. Filter evaluations

We included in the review evaluations that fulfilled the requirements described in Table 9.

Table 9: Inclusion criteria

Requirements
1. The Intervention must have been evaluated in multiple sites or at different points of time
2. The effects were assessed using experimental methods
3. The evaluation was published after 1990
4. The intervention was a parenting program for children 0-5 years old
5. The reported outcomes included at least cognitive or behavioral skills
6. The evaluation compared the treatment group to the status quo
7. There were sufficient information to compute effect sizes
8. The evaluations including components unrelated to parenting (e.g., nutrition) were excluded

Because evaluations tend to report multiple effects, we followed these criteria to determine the effects to record:

(i) Only effects for the whole population were included (i.e., effects for sub-populations were excluded). When only effects for sub-samples were reported (e.g., by age or cohort), then we extract these effects.
(ii) If evaluation reported ITT and TOT effects, we gave priority to ITT.
(iii) The effects from clustered randomized control trials were included, even when the standard errors were not adjusted for clustering.
(iv) Only endline effects were included. If not available, effects of the first available follow-up after the endline were included.
(v) Only effects for behavioral and cognitive outcomes measured with valid psychometric tests were considered.
(vi) To compute effect sizes we gave priority to formulas based on means. If not available, we used regression coefficients.
(vii) Outlier effects (larger than 2 SD in absolute value) were excluded.

2.3 Code variables

No departures.

2.4 Categorize interventions

Interventions reporting effects in cognitive skills were categorized in the program types described in Table 10. Note that because we put a requirement that interventions had to be evaluated in multiple sites, the included interventions correspond to specific and typically well-known programs. For example, we included evaluations of Early Head Start, the public program implemented in the United States, as well as other well-known programs, such as Nurse Family Partnership.

Table 10: Program types (outcome: cognition)

Program types	Description
Early Head Start	US governmental program providing daycare, home visits or combination of both
Healthy Families America	Weekly, then less frequent home visits
HIPPY*	Biweekly home visits and parent group meetings with paraprofessional
Infant Health and Development Program	Daycare, home visits and parent group meetings for low birthweight 0-3 year olds
Jamaica curriculum	Weekly home visits to promote age-appropriate stimulation activities
Nurse Family Partnership	Home visits for mothers from pregnancy until child is 2 years old by a nurse
Parents as Teachers	Monthly home visits and group meetings

Notes: HIPPY stands for Home Instruction Program for Preschooler Youngsters

The program types that were evaluated in terms of improving child behavior are presented in Table 11.

Table 11: Program types (outcome: behavior)

Program types	Description
Early Head Start	US governmental program providing daycare, home visits or combination of both
Family Check-Up	Individual visits for at-risk children with therapist
Healthy Families America	Weekly, then less frequent home visits
Healthy Steps	Pediatric office visits and home visits by nurse or trained worker
HIPPY*	Biweekly home visits and parent group meetings with paraprofessional
Incredible Years	Weekly group sessions for parents
Infant Health and Development Program	Daycare, home visits and parent group meetings for low birthweight 0-3 year olds
Nurse Family Partnership	Home visits for mothers from pregnancy until child is 2 years old by a nurse
Parent-Child Interaction Therapy	Weekly 1-hour group sessions for parents
Parents as Teachers	Monthly home visits and group meetings
Positive Parenting Program	Visits to practitioner, home visits or self-study

Notes: HIPPY stands for Home Instruction Program for Preschooler Youngsters

2.5 Compute effect sizes

In this review, the procedure for computing effect sizes was not exactly the same as the procedure followed in the benchmark review. We explored whether the results were robust to employing the approach aligned with the benchmark, and we have found that the differences were minor. Still, in the future revision of this review, we will evaluate the possibility of following the benchmark procedure for computing effect sizes.

In particular, in this review the effect sizes were computed using the raw means and standard deviations pre, post, for treatment and control groups whenever possible. If pre-treatment scores were available, then the effect size was computed as gains in difference between treatment and control means divided by the pooled standard deviation:

$$Effect\ Size = \frac{Y_{T\ post} - Y_{C\ post}}{S_{pooled\ post}} - \frac{Y_{T\ pre} - Y_{C\ pre}}{S_{pooled\ pre}}$$

where

$$S_{pooled\ post} = \sqrt{\frac{\left(N_{T\ post} - 1\right)SD_{T\ post}^2 + \left(N_{C\ post} - 1\right)SD_{C\ post}^2}{N_{T\ post} + N_{C\ post} - 2}}$$

and

$$S_{pooled\ pre} = \sqrt{\frac{(N_{T\ pre} - 1)SD^2_{T\ pre} + (N_{C\ pre} - 1)SD^2_{C\ pre}}{N_{T\ pre} + N_{C\ pre} - 2}}$$

and $Y_{T\ post}$: treatment group mean post treatment; $Y_{T\ pre}$: treatment group mean before treatment; $Y_{C\ post}$: control group mean post treatment; $Y_{C\ pre}$: control group mean before treatment; $SD_{T\ post}$: treatment group standard deviation post treatment; $SD_{T\ pre}$: treatment group standard deviation before treatment; $SD_{C\ post}$: control group standard deviation post treatment; $SD_{C\ pre}$: control group standard deviation before treatment; $N_{T\ post}$: treatment group number of observations post treatment; $N_{T\ pre}$: treatment group number of observations before treatment; $N_{C\ post}$: control group number of observations post treatment; $N_{C\ pre}$: control group number of observations before treatment.

If pre-treatment scores were not available, then the effect sizes were computed as post-treatment differences in means between treatment and control groups divided by post-treatment pooled standard deviation:

$$Effect\ Size = \frac{Y_{T\ post} - Y_{C\ post}}{S_{pooled\ post}}$$

where

$$S_{pooled\ post} = \sqrt{\frac{(N_{T\ post} - 1)SD^2_{T\ post} + (N_{C\ post} - 1)SD^2_{C\ post}}{N_{T\ post} + N_{C\ post} - 2}}$$

and $Y_{T\ post}$: treatment group mean post treatment; $Y_{C\ post}$: control group mean post treatment; $SD_{T\ post}$: treatment group standard deviation post treatment; $SD_{C\ post}$: control group standard deviation post treatment; $N_{T\ post}$: treatment group number of observations post treatment; $N_{C\ post}$: control group number of observations post treatment.

When there was not enough information to compute the effect size using raw means, we used the regression coefficient reported in the paper.

Finally, standard errors of all effect sizes were computed using the following formula:

$$Standard\ Error = \sqrt{\frac{N_{T\ post} + N_{C\ post}}{N_{T\ post}N_{C\ post}} + \frac{Effect\ size^2}{2(N_{T\ post} + N_{C\ post})}}$$

where $N_{T\ post}$: treatment group number of observations post treatment; $N_{C\ post}$: control group number of observations post treatment.

2.6 Combine effect sizes

No departures.

2.7 Generate results

No departures.

3 Secondary Enrollment and Completion review

This section presents the departures from the benchmark methodology that occurred when producing the Secondary Enrolment and Completion review.

3.1. Search papers

We collected references from recent reviews in the literature reported in Table 12. Additionally, we performed keyword searches using Google Scholar and selected journals.

Table 12: Existing reviews

References
Fryer (2016)
Vivalt (2016)
Stevenson et al. (2015)
McEwan (2015)
Glewwe and Muralidharan (2015)
Lavecchia et al. (2014)
Murnane and Ganimian (2014)
Cullen et al. (2013)
Petrosino et al. (2012)

3.2. Filter evaluations

We included in the review evaluations that fulfilled the requirements described in Table 13.

Table 13: Inclusion criteria

Requirements
1. The intervention was carried out at the secondary level or the outcome was captured at the secondary level
2. The outcomes of the intervention included some measure of coverage (enrollment, dropout) or completion (grade progression, graduation from level)
3. The treatment group was compared to the status quo
4. The effects were estimated using RCT, RD, IV, differences-in-differences or propensity score matching
5. The effects were based on dichotomous outcomes or proportions (e.g., share of school aged children enrolled)
6. The estimates were calculated from a linear probability model or Probit model, or could be interpreted as a percentage point change in the probability of an outcome occurring
7. There were sufficient information to compute effect sizes

Because evaluations tend to report multiple effects, we followed these criteria to determine the effects to record:

(i) Only effects for the whole population were included (i.e., effects for sub-populations were excluded). When only effects for sub-samples were reported (e.g., by age, gender, geographic area), then we extract all these effects.
(ii) If evaluation reported ITT and TOT effects, we gave priority to ITT.
(iii) The effects from clustered randomized control trials were included, even when the standard errors were not adjusted for clustering.
(iv) The effect for the first follow-up was extracted.
(v) If an evaluation reported effects from different research designs, we extracted the one preferred by the author and if there were no one that was suggested as preferred, we used the ITT specification.
(vi) Effects for only one specification were extracted following pre-established criteria.
(vii) Outlier effects (larger than 15 percentage points improvement) were excluded.

3.3. Code variables

No departures.

3.4. Categorize interventions

The interventions reporting effects on enrollment were categorized in the program types described in Table 14. The interventions reporting effects on completion were categorized in the program types described in Table 15.

Table 14: Program types (outcome: enrolment)

Program types	Description
Conditional cash transfers	Cash transfers to households conditional on school enrolment
Counseling, coaching and information	Information on returns to schooling, counseling on academic choices
Scholarships and achievement awards	Cash to cover school fees, awards for good performance at school
School inputs	Teacher training, textbooks, management practices, infrastructure improvements
Unconditional cash transfers	Cash transfers to households unconditional on school enrolment

Table 15: Program types (outcome: completion)

Program types	Description
Conditional cash transfers	Cash transfers to households conditional on school enrollment
Counseling, coaching and information	Information on returns to schooling, counseling on academic choices
Scholarships and achievement awards	Cash to cover school fees, awards for good performance at school
School inputs	Teacher training, textbooks, management practices, infrastructure improvements
Packaged interventions	Integral strategies with more than one type of intervention bundled together

3.5. Compute effect sizes

The effect size used in the analysis was the raw standardized differences in terms of percentage point increases in enrollment, drop out, completion (i.e., graduation from secondary school) and adequate progression to the following grade. To reduce the number of involved outcomes we combined the effects on enrollment with the effects on drop out. Note that the effects on drop out were multiplied by -1 for consistency (i.e., reductions in drop out should correspond to increases in enrolment). Additionally, we combined the effects in completion with the effects in adequate progression to the following grade as these two measures are linked to successful completion of secondary education. Because the effect sizes were expressed in terms of raw standardized differences, we could extract effects (and standard errors) reported in percentage points ready for the analysis.

3.6. Combine effect sizes

No departures.

3.7. Generate results

No departures.

4. Learning in Secondary review

This section presents the departures from the benchmark methodology that occurred when producing the Learning in Secondary review.

4.1. Search papers

We collected references from recent reviews in the literature reported in Table 16.

Table 16: Existing reviews

References
Baird et al. (2013)
Banerjee et al. (2013)
Cheung and Slavin (2013)
Fryer (2016)
Glewwe et al. (2011)
Kremer and Holla (2009)
Krishnaratne et al. (2013)
Murnane and Ganimian (2014)
Rakes et al. (2010)
Slavin et al. (2008)
Slavin et al. (2009)
Snilstveit et al. (2015)

4.2. Filter evaluations

No departures.

4.3. Code variables

No departures.

4.4. Categorize interventions

The interventions assessed in the evaluations included in the review were categorized in the program types described in Table 17.

Table 17: Program types

Program types	Description
Pedagogical practices	Introduction of new practices, instructional material, training or coaching
Monetary incentives to teachers	Monetary incentives to teachers based on performance
Competitive teacher selection	Introduction of a competitive selection process to become a teacher
Alternative teacher selection	Introduction of alternative pathways to become a teacher
Extended school day	Extending the school day
Technology	Use of technology for instruction
Cash transfers	Cash transfers to households
Vouchers, subsidies or scholarships	Vouchers to send children to the school of their choice
Monetary incentives to students	Monetary incentives to students based on performance
Information on returns	Information on returns to schooling and requirements for college
No excuses charter schools	Comprehensive programs that are offered in No Excuses charter schools

4.5. Compute effect sizes

No departures.

4.6. Combine effect sizes

No departures.

4.7. Generate results

No departures.

References

Baird, Sarah, Francisco HG Ferreira, Berk Özler, and Michael Woolcock. "Relative effectiveness of conditional and unconditional cash transfers for schooling outcomes in developing countries: a systematic review." Campbell systematic reviews 9, no. 8 (2013).

Banerjee, Abhijit, Paul Glewwe, Shawn Powers, and Melanie Wasserman. "Expanding access and increasing student learning in post-primary education in developing countries: A review of the evidence." J-PAL, 2013.

Borenstein, Michael, Larry Hedges, and Julian Higgins, Hannay Rothstein, 2009. "Introduction to Meta-Analysis." Chichester, UK, John Wiley & Sons, Ltd. doi 10: 9780470743386.

Cheung, Alan CK, and Robert E. Slavin. "The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis." Educational research review 9 (2013): 88-113.

Cullen, Julie Berry, Steven D. Levitt, Erin Robertson, and Sally Sadoff. "What can be done to improve struggling high schools?" The Journal of Economic Perspectives 27, no. 2 (2013): 133-152.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. "Incentives work: Getting teachers to come to school." The American Economic Review 102, no. 4 (2012): 1241-1278.

Fryer Jr, Roland G. The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. No. w22130. National Bureau of Economic Research, 2016.

Glewwe, Paul, and Karthik Muralidharan. "Improving school education outcomes in developing countries: evidence, knowledge gaps, and policy implications." University of Oxford, Research on Improving Systems of Education (RISE). Working paper RISE-WP-15/001 (2015)

Glewwe, Paul, and Karthik Muralidharan, Chapter 10 - Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications, In: Eric A. Hanushek, Stephen Machin and Ludger Woessmann, Editor(s), Handbook of the Economics of Education, Elsevier, 2016, Volume 5, Pages 653-743, ISSN 1574-0692, ISBN 9780444634597

Glewwe, Paul W., Eric A. Hanushek, Sarah D. Humpage, and Renato Ravina. School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. No. w17554. National Bureau of Economic Research, 2011.

Kremer, Michael, and Alaka Holla. "Improving education in the developing world: What have we learned from randomized evaluations?" Annu. Rev. Econ. 1, no. 1 (2009): 513-542.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. "The challenge of education and learning in the developing world." Science 340, no. 6130 (2013): 297-300.

Krishnaratne, Shari, Howard White, and Ella Carpenter. "Quality education for all children? What works in education in developing countries." New Delhi: International Initiative for Impact Evaluation (3ie), Working Paper 20 (2013).

Lavecchia, Adam M., Heidi Liu, and Philip Oreopoulos. Behavioral economics of education: Progress and possibilities. No. w20609. National Bureau of Economic Research, 2014.

McEwan, Patrick J. "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments." Review of Educational Research 85, no. 3 (2015): 353-394.

Muralidharan, Karthik, and Venkatesh Sundararaman. "Teacher performance pay: Experimental evidence from India." Journal of political Economy 119, no. 1 (2011): 39-77.

Murnane, Richard J., and Alejandro J. Ganimian. Improving educational outcomes in developing countries: Lessons from rigorous evaluations. No. w20248. National Bureau of Economic Research, 2014.

Petrosino, Anthony, Claire Morgan, Trevor Fronius, Emily Tanner-Smith, and Robert Boruch. "Interventions in developing nations for improving primary and secondary school enrollment of children: A systematic review." Campbell Systematic Reviews 8, no. 19 (2012).

Rakes, Christopher R., Jeffrey C. Valentine, Maggie B. McGatha, and Robert N. Ronau. "Methods of instructional improvement in algebra: A systematic review and meta-analysis." Review of Educational Research 80, no. 3 (2010): 372-400.

Slavin, Robert E., and Cynthia Lake. "Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis. Version 1.2." Best Evidence Encyclopedia (BEE), 2007.

Slavin, Robert E., Alan Cheung, Cynthia Groff, and Cynthia Lake. "Effective reading programs for middle and high schools: A best-evidence synthesis." Reading Research Quarterly 43, no. 3 (2008): 290-322.

Slavin, Robert E., Cynthia Lake, and Cynthia Groff. "Effective programs in middle and high school mathematics: A best-evidence synthesis." Review of Educational Research 79, no. 2 (2009): 839-911.

Snilstveit, Birte, Emma Gallagher, Daniel Phillips, Martina Vojtkova, John Eyers, Dafni Skaldiou, Jennifer Stevenson, Ami Bhavsar, and Philip Davies. Education Interventions for Improving the Access to, and Quality of, Education in Low and Middle Income Countries: A Systematic Review. International Initiative for Impact Evaluation (3ie), 2015.

Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M.G. and J. Eyers. Interventions for improving learning outcomes and access to education in low-and middle-income countries: a systematic review. International Initiative for Impact Evaluation (3ie), December 2015.

Vivalt, Eva. "How Much Can We Generalize from Impact Evaluations?" Mimeo, Stanford University (March 28, 2016).

Worth, Jack, Juliet Sizmur, Rob Ager and Ben Styles. "Improving Numeracy and Literacy Evaluation report and Executive summary." Education Endowment Foundation (2015).