Impossible generalisations: meta-analyses of education interventions in international development

Edoardo Masset*

June 11, 2019

Abstract

Effectiveness reviews of education interventions in low and middle income countries have found divergent results and have struggled to make sense of heterogeneous evidence. Researchers have employed the right random-effects models in meta-analyses, but the results of these meta-analyses are often misunderstood. In this paper we calculate prediction intervals for several meta-analyses of education interventions. Unlike confidence intervals, prediction intervals take into account differences between studies in addition to sampling variation. The results show that the impact of education interventions is unpredictable. No intervention appears to be more effective than others for any of four key outcomes. Given the high heterogeneity of results, impacts of education interventions cannot be generalised, and researchers should focus on investigating the sources of this heterogeneity rather than trying to identify which interventions work.

*Center of Excellence in Development Impact and Learning (CEDIL) and London School of Hygiene and Tropical Medicine, LIDC 20 Bloomsbury Sqare, London WC1A 2NS. Email: edoardo.masset@lshtm.ac.uk

1 Introduction

Evidence-based policy presumes an understanding of 'what works', but what works in a specific context is rarely known. More often, policy decision are based on 'what worked' in other places or at other times. Evidence-based policy therefore relies on the generalisation of evidence from one context to another. Several approaches to generalising evidence are possible, and meta-analysis is a very popular one.

Meta-analyses synthesise quantitative results of studies and summarise bodies of knowledge. Meta-analyses are common in education, and indeed some of the very first meta-analyses summarised the results of education studies (Shadish et al., 2002). More recently, meta-analyses have been used to summarise the evidence on the impact of education interventions in international development. Given the ever increasing size of evidence from impact evaluations, and given the variety of interventions and contexts, the researchers' aspiration to summarise quantitative results through meta-analyses is understandable. However, there are some misconceptions about what meta-analyses do and about how their results should be interpreted. A common misconception among critics of meta-analyses is that they aim at finding a universal effect applicable to all contexts (Deaton and Cartwright, 2018). Unfortunately, the same misconception is often shared by researchers who hope to use meta-analyses to identify 'overall effects' or the 'most effective' or 'most promising' interventions.

Several meta-analyses of education interventions in low and middle income countries have been recently conducted. These meta-analyses have found conflicting results and struggled to make sense of an heterogeneous body of evidence. In this paper we show that when the results of meta-analyses of education interventions are correctly interpreted, no generalisations are possible. No education intervention can be expected to be more effective than others, and every intervention can produce a wide range of results, including no impact. We argue that attempts to generalise evidence on education interventions and to estimating 'overall effects' are vain and potentially misleading. Quantitative reviews of evidence should rather focus their attention on understanding the sources of heterogeneity and on explaining how impacts vary as contexts and characteristics of implementation change.

2 Fixed and random-effects meta-analyses

Two types of meta-analysis are common: random-effects meta-analyses and fixed-effect meta-analyses.[1] In a fixed-effect meta-analysis, we assume that the studies analyse a common effect. The researcher collects data on interventions whose results vary only because of sampling variation. Different studies find different effects drawing from a distribution of possible effects around the 'true' effect. Large studies will be more likely to be close to the 'true' effect, while small studies will be more likely to be off. As the number of studies increases,

[1]The latter not to be confused with regression models of panel data with time-invariant individual slopes

the estimated average effect gets closer to the true effect. Implicit to the assumption of a common effect is the assumption of a universal effect that generalises to all contexts. Once the common effect has been estimated by a sufficiently large number of studies, the results of a fixed-effect meta-analysis apply always and everywhere.

While the assumption of a common effect is plausible in some cases, in most cases we expect the effects estimated by different studies to be different for a number of reasons. Sources of heterogeneity are many and include: differences in characteristics across populations, differences in implementation of the interventions, differences in outcomes or how outcomes are measured, differences in exposure to treatments, and various biases produced by the quality of study design and study implementation. One of the main goals of meta-analysis is to produce a summary measure of the evidence to conclude, for example, that structured pedagogy interventions improve, on average, mathematics test score by 0.2 standard deviations. However, it is difficult to decide what to make of this average effect when the heterogeneity of effects is very large.

If we abandon the belief in the existence of a common effect and recognise that different effects are expected depending on various characteristics of implementation, then a random-effects model is preferable to a fixed-effect model. A random-effects meta-analysis assumes that effects from different studies, though similar, are different. Hence, it is assumed that different effects observed by different studies include both a within-study variation, due to sampling error, and a between-study variation, due to the differences between studies. The assumption of a common effect is untenable in most cases, and certainly for studies of development interventions, which are highly heterogeneous in populations, interventions, and study types. It should be no surprise therefore that random-effects meta-analyses have become the norm.

Results of random-effects meta-analyses however are often misunderstood. In particular, the idea that a random-effects meta-analysis identifies an 'overall effect' is a common misconception (Higgins et al., 2009). The average effect of a random-effects meta-analysis bears limited information. Given the diversity of the studies and interventions, by no means it implies that on average this is the result that would be observed by a new study. The summary effect of a random-effects meta-analysis is not a generalisable effect, it is simply the mean of a distribution of effects. In a fixed-effect meta-analysis the summary effect is the true effect size of the intervention. In a random-effects meta-analysis, we assume that the true effect size varies from study to study and therefore the summary effect is simply an average of these different effects (Borenstein et al., 2009).

A second misconception of random-effects meta-analysis concerns the confidence interval, and the belief that confidence intervals represent the heterogeneity across studies (Higgins et al., 2009). Confidence intervals of random-effects meta-analyses are wider than confidence intervals of fixed-effect meta-analysis, but this is not a reflection of the uncertainty related to the variety of effects. Confidence intervals of random-effects meta-analysis do not represent

the likelihood of occurrence of the effect. The preferred measure to assess the likelihood of occurrence of an effect is the prediction interval. Prediction intervals define the range of effects to be expected from a new intervention which is similar to those included in the meta-analysis. Prediction intervals, rather than confidence intervals, should be used to interpret the generalisability of the results of a meta-analysis. A bit of algebra will help clarifying.

In a fixed-effect meta-analysis, the summary effect is a weighted average of the effects estimated in each study, where the weights (w_i) are the reciprocal of the within-study variance ($\sigma^2_{w_i}$):

$$w_i = \frac{1}{\sigma^2_{w_i}} \tag{2.1}$$

The variance of the mean effect (μ) is:

$$\sigma^2_{\mu} = \frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma^2_{w_i}}} \tag{2.2}$$

and 95% confidence intervals for the mean effect are:

$$CI = \mu \pm Z^{\alpha} \sqrt{\sigma^2_{\mu}} \tag{2.3}$$

In a random-effects meta-analysis, the weight applied to each study includes not only the within-study variance, but also an estimate of the between-study variance (σ^2_b):

$$w_i = \frac{1}{\sigma^2_{w_i} + \sigma^2_b} \tag{2.4}$$

and the variance of the mean effect is:

$$\sigma^2_{\mu} = \frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma^2_{w_i} + \sigma^2_b}} \tag{2.5}$$

Given the weights in 2.4, studies with small within-study variance have more weight, but they weight less than they would in a fixed-effect model.Conversely, studies with large within-study variance weight less, but more than they would in a fixed-effect meta-analysis. The inclusion of the between-study variance in the calculation of the variance of the average effect produces a shrinkage of the relative importance of each study.

The confidence intervals of a random-effects model are calculated analogously to those of a fixed-effect model in 2.3, but the variance of the mean effect now includes the within-study variance as well as the between-study variance. As a result, the confidence intervals of a random-effects meta-analysis are wider than those of a fixed-effect meta-analysis. However, it is important to note that as the number of studies increases, the variance of the mean effect decreases regardless of the between-study variance. As the number of included studies increases, the confidence intervals become narrower unless the heterogeneity between studies

also increases. In other words, the confidence intervals of a random-effects meta-analysis do not reflect the heterogeneity of the included studies.

A measure of heterogeneity in a random-effects meta-analysis is provided by prediction intervals. The concept of a prediction interval is familiar. Given a mean and a standard deviation for a variable drawn from a normal distribution, we expect the value of the variable to fall within the interval $\mu \pm Z^{\alpha}\sigma_{\mu}$, 95% of the times. In meta-analysis, the prediction interval is analogously built and allows for both the uncertainty resulting from within study sampling error and for the uncertainty resulting from between-study heterogeneity. A formula for calculating prediction intervals is (Borenstein et al., 2009):

$$PI = \mu \pm t^{\alpha}_{df}\sqrt{\sigma^2_{\mu} + \sigma^2_{b}} \tag{2.6}$$

This formula uses the variance of the mean effect and the estimated between-study variance to calculate the standard deviation. The standard deviation is then multiplied by the value corresponding to the chosen significance level α of a t distribution with $k-2$ degrees of freedom (df), where k is the number of studies included in the meta-analysis.

Confidence intervals and prediction intervals have very different meaning. A confidence interval says that in 95% of cases the mean effect falls within the interval. The prediction interval says that in 95% of cases a new study, which is similar to the included studies, will fall within the interval. The confidence interval simply informs about the accuracy in the estimation of the mean effect, while the prediction interval informs about the dispersion of effects around the mean. If the prediction interval does not include zero, then we are confident that in 95% of cases a new study will estimate an effect larger than zero.

It is important to note that prediction intervals apply to studies that are 'similar' to those already included. The prediction interval does not provide a good prediction for studies that are very different from those included in the meta-analysis. It is also important to note that studies included in a meta-analysis are not drawn randomly from a normal distribution. Because of publication bias and because of the process of scientific research, studies included in a meta-analysis tend to be more homogeneous than they really are. For example, researchers tend to find results that are similar to those of previous studies, or they use similar study designs, or conduct studies in the same countries or contexts. As a result, the prediction intervals are narrower than what they would be if studies were randomly drawn from a population of studies. For these two reasons: applicability of prediction intervals to similar studies and artificial homogeneity produced by publication bias and the research process, prediction intervals are conservative estimates of the true heterogeneity of effects. The heterogeneity of effects is likely to be larger than the heterogeneity described by the prediction intervals, thus limiting their predictive ability.

One limitation of prediction intervals is that they may capture the heterogeneity in the quality of the primary studies in addition to the heterogeneity of effects (Riley et al., 2011).

If primary studies are biased because of poor design, for example because the study arms were not randomised or because the sample sizes were small, the biases will be included in the calculation of the prediction intervals (Higgins and Green, 2011). Prediction intervals therefore are more appropriate when the studies included in the meta-analysis are of high-quality and have limited bias, something that can be investigated by an analysis of risk of bias and of publication bias.

Given that the main goal of a random-effects meta-analysis is understanding the heterogeneity of effects and predicting the effect of future interventions, one would expect a wide use of prediction intervals in the practice of meta-analysis. Yet most researchers report confidence intervals of mean effects rather than prediction intervals, worryingly allowing the inexperienced reader to interpret these confidence intervals as the uncertainty about whether a new intervention will work or not.[2] In a recent review (IntHout et al., 2016) it was found that in more than 70% of statistically significant random-effects meta-analysis from the Cochrane database, the prediction intervals would have included the zero value, suggesting that there were contexts in which a new similar intervention would not have been effective. This is a very different message from the one suggested by a confidence interval of the effect size that does not include zero. Though the confidence interval of a random-effects meta-analysis simply represents the accuracy of the estimated mean, it is often interpreted as the likely range of effects of the intervention, and a mean effect whose confidence interval does not include zero is often interpreted as meaning that the 'overall effect' is statistically different from zero.

In the next section we compare confidence intervals and prediction intervals of a series of meta-analyses of education interventions, showing how their use leads to very different conclusions.

3 Meta-analyses of education studies in international development

There is a long tradition of meta-analyses in education studies and indeed one of the very first meta-analyses ever conducted calculated the mean effect of class size on students achievements (Glass and Lee, 1979). This first meta-analysis did not hide the goal of generalising evidence by identifying a common effect, and concluded that 'there is little doubt that, other things equal, more is learnt in small classes.' This first attempt at summarising education studies was followed by many others, as meta-analysis techniques gained acceptance and popularity. More recently there have been several attempts to summarise the evidence on what works in education in international development, which is the focus of the present

[2]The study by Lawry et al. (2014) is a rare example of a random-effects meta-analysis in international development that employs prediction intervals rather than confidence intervals.

paper.

Evans and Popova (2016) reviewed six systematic reviews of education interventions which sought to improve learning outcomes in developing countries, three of which included meta-analyses (Conn, 2014; Krishnaratne et al., 2013; McEwan, 2015). Both McEwan (2015) and Krishnaratne et al. (2013) presented the results of random-effects meta-analyses. In the papers, charts show mean effects with confidence intervals as it is customary in the literature. Interestingly, the six reviews, including the three meta-analyses, produced diverging results. Conn (2014) found that pedagogical interventions and students incentives were the most effective. Krishnaratne et al. (2013) identified the provision of material as most effective, while McEwan (2015) found several effective interventions, such as computers and instructional technology, teacher training, smaller classes, smaller learning groups within classes, contract and volunteer teachers, students and teacher performance incentives, and instructional material.

Evans and Popova (2016) set out to investigate the reasons for the divergence of conclusions of these reviews. They concluded that the differences were partly the result of how interventions were categorised, and partly the result of the primary studies considered. In relation to the first point, there are no agreed standards on how different interventions should be categorised, and reviewers employ a good deal of subjectivity in building intervention categories. In addition, interventions often include different components and can end up in different categories. For example, Krishnaratne et al. (2013) categorised as 'material', interventions that were defined by others as 'computer-aided learning' or 'teaching incentives'. Also confusingly, there was significant overlap between studies defined by Conn as 'pedagogical interventions' and studies classified by McEwan as 'computers or instructional technology.'

However, much of the difference in the conclusions was driven by the lack of overlap between the primary studies reviewed. All reviews adopted different criteria for searching, identifying and selecting the included studies. The meta-analyses summarised the results of different studies, and it should not be a surprise that they reached different conclusions. As a result, Evans and Popova (2016) recommended that reviewers should perform more exhaustive searches of the literature. Another interpretation of the difference in the conclusions is that the available evidence is highly heterogeneous. If the results were highly homogeneous, the difference in overlap between reviews would not matter and the mean effects would be similar. On the other hand, if the results reported by the primary studies are highly heterogeneous, then the summaries of results produced by these meta-analyses are likely to be also different. If heterogeneity between studies is large, increasing the number of studies alone does not help generalisations. This is confirmed by another systematic review by Snilstveit et al. (2015).

The study by Snilstveit et al. (2015) is the most comprehensive systematic review of education effectiveness studies in developing countries conducted to date. It includes results from

238 experimental and quasi-experimental studies assessing 216 different programmes. The interventions were classified in 20 different categories at different levels: child level (providing information, merit-based scholarship, school-based health, and school feeding), household level (cash transfers and reducing fees), school level (remedial education, new schools and infrastructure, providing materials, structured pedagogy, grouping by ability, extra time, and computer-assisted learning), teacher level (teacher training, hiring teachers, and teachers incentives), system level (public-private partnerships, school-based management, and and community-based monitoring), and multi-component interventions.

The authors of the review conducted random-effects meta-analyses for each category separately and presented summary effects together with their confidence intervals. They considered participation outcomes (enrolment, attendance, drop-out, and completion) and learning outcomes, such as scores on mathematics and literacy tests. They also reported standard measures of heterogeneity, in particular the between-study variance within each category, which we will use in this paper to calculate prediction intervals. The review, whose results were also summarised in Snilstveit et al. (2016), concluded that conditional cash transfer interventions have the most substantial effect on school participation, while community-based monitoring, low-cost private schools, new schools and infrastructure, and school feeding are 'promising' interventions. Further, according to the review, structured pedagogy interventions have the largest impact on learning, while merit-based scholarship, school-feeding, extra-time and remedial education are also 'promising.' For some other intervention categories, the authors found null or negative effects, i.e. summary effects whose confidence intervals included zero.

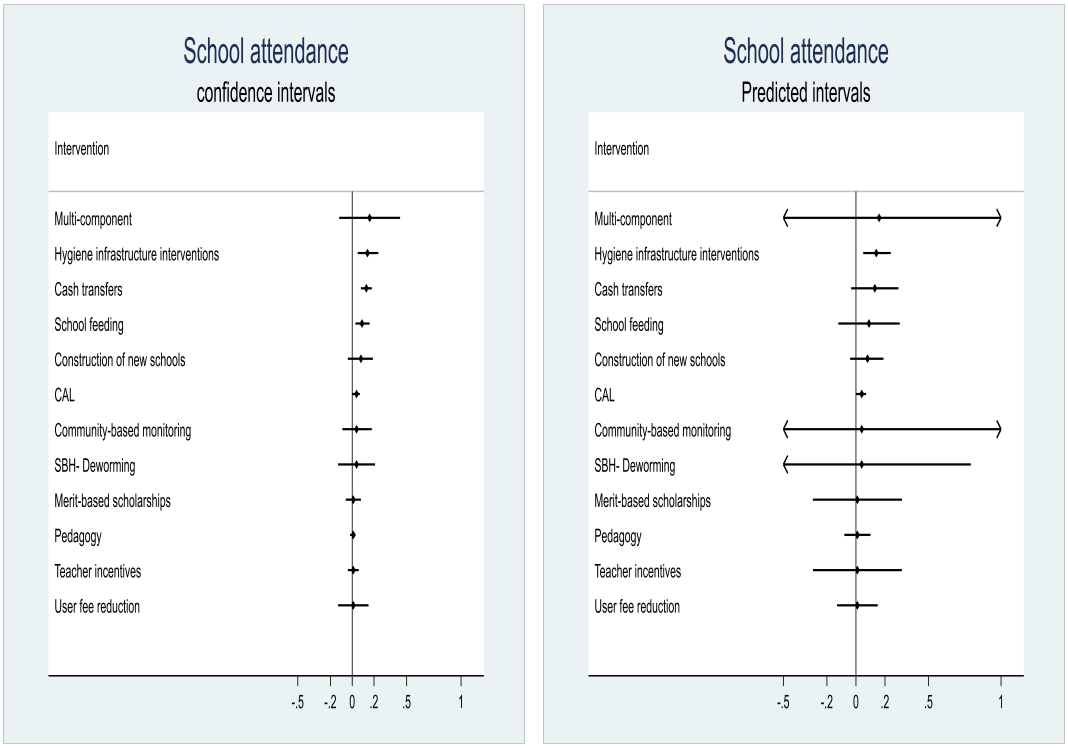
In this paper we used the same data used by Snilstveit et al. (2015) to calculate prediction intervals for each intervention category. We then compared the results obtained using confidence intervals to those obtained using prediction intervals. We conducted this exercise for two key participation outcomes (school attendance and school completion) and for two key learning outcomes (mathematics and literacy test scores). The original meta-analyses in Snilstveit et al. (2015) were conducted after transforming the original outcomes in standardised effect sizes. The outcomes were divided by the sample standard deviation in each primary study in order to be expressed in the same currency. The practice is standard in the calculation of test scores and the authors followed the same approach for effect sizes of all binary outcomes such as, for example, school attendance. The practice of standardising outcomes has a number of problems of its own (Simpson, 2017) but we will not discuss them here.

We present the results of our calculations in the figures below. In each figure, the chart on the left shows the summary effects of random-effects meta-analyses with their confidence intervals, while the chart on the right shows the summary effects and the prediction intervals. In all charts, each row represent a meta-analysis for each intervention category - not a primary study. The interventions are ordered in descending order of the mean effect size.

As usual, points to the left of the zero line show a beneficial impact of the intervention and a confidence interval not crossing the zero line gives us confidence on the strength of the result. For example, in the left chart of figure 1, hygiene infrastructure interventions, cash transfers and school feeding, show statistically significant effects. Other interventions have moderately large confidence intervals but nevertheless fail the statistical significance test. This evidence led Snilstveit et al. (2015) to conclude that conditional cash transfer are effective in increasing attendance rates.

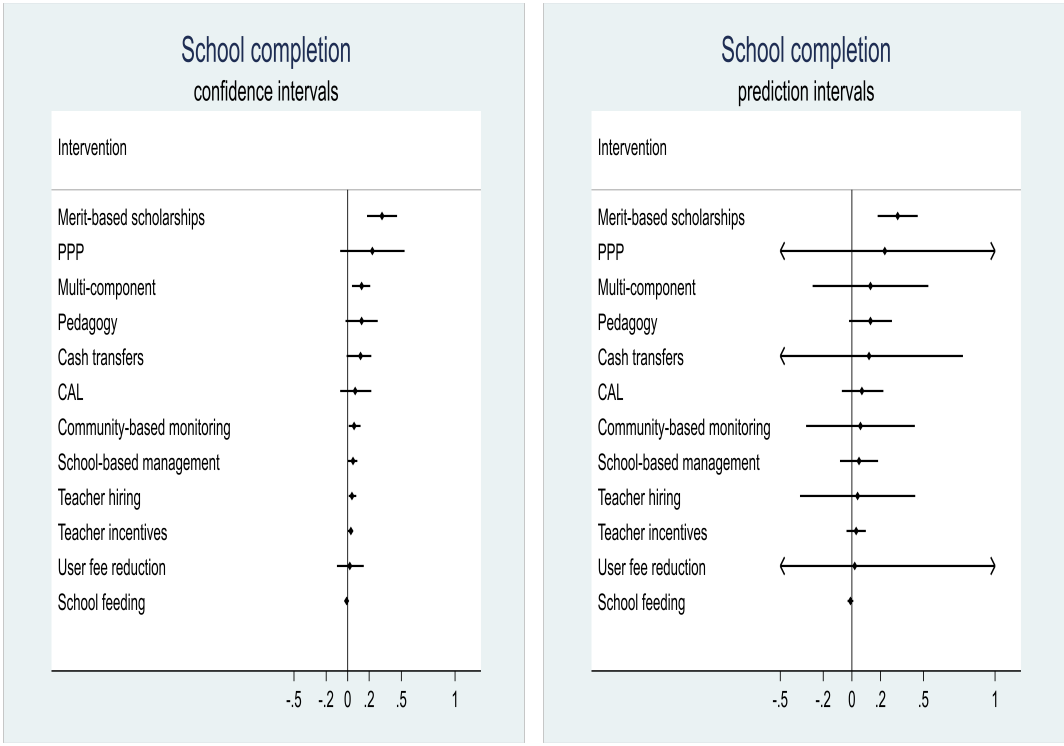
The chart on the right hand side of figure 1 shows the same summary effects using prediction intervals. Recall that prediction intervals represent the expected impact of a new study, which is similar to the studies included in the review. In the right chart of figure 1, all prediction intervals are much wider than the confidence intervals. They show a large range of effects for each intervention, and only one prediction interval (hygiene infrastructure interventions) does not cross the zero line. Based on the right chart of figure 1, it is difficult to say which intervention is most effective as none, including cash transfers, is statistically different from zero. Note that the positive impact found for hygiene infrastructure intervention should not be overestimated because the meta-analysis for this intervention was based on just two primary studies. Prediction intervals can only be calculated when the meta-analysis includes at least three studies. In this case, as in the following charts, the between-study variance was set to zero when there were fewer than three studies, so that confidence intervals and prediction intervals are identical. More in general, when the number of studies is very small, the between-study variance cannot be estimated with precision and prediction intervals should be used with caution. Viable alternative approaches in these cases are the use of plausible values for the between-study variance (Higgins et al., 2009) or the presentation of the results of the individual studies rather than the overall summary effect (Cox and Keogh, 2015).

Figure 1: Meta-analyses of impacts on attendance rates



We found similar results for meta-analyses of completion rates (see figure 2). Prediction intervals are much wider than confidence intervals. They show a wider range of results, which often include large negative and positive impacts. The arrows at the edges of some of the prediction intervals mean that the interval is actually wider than the fixed range represented in the chart. Only one intervention (merit-based scholarship) shows an impact significantly different from zero. As in the case of school attendance, it is very hard to conclude that cash transfers unequivocally improve completion rates, and it is not clear in what way other interventions are 'promising.' It seems that all the interventions may or may not work depending on the context or on the study considered. Some interventions show a larger heterogeneity than others, which would be an interesting result if it reflected the true dispersion of effect sizes for each intervention. We would be able to say, for example, that impacts of structured pedagogy are more predictable than those of cash transfers. However, it is difficult to say to what extent the size of the prediction interval is driven by the heterogeneity of the effects or by the lack of precision in measuring the between-study variance, which is highly imprecise when the meta-analysis includes only few studies.

Figure 2: Meta-analyses of impacts on completion rates



Figures 3 and 4 show meta-analyses of learning outcomes. There were more primary studies on learning outcomes and therefore we were able to conduct our meta-analyses for a larger number of intervention categories. Based on the confidence intervals, structured pedagogy and merit-based scholarship have a statistically significant impact on mathematics test scores. However, the prediction intervals of pedagogy and scholarship are quite large and not significantly different from zero. In fact their prediction intervals are comparable in size to those of other interventions and they do not appear to be more effective than other interventions. Based on prediction intervals, none of the interventions excludes zero among the potential outcomes.

Impacts of the interventions on reading test scores present a similar pattern. Prediction intervals are much wider than the confidence intervals. Only one intervention type (extra time) shows a prediction interval different from zero, but this result is based on only two studies. The between-study variance for this intervention was set to zero, in such a way that the confidence intervals and the prediction intervals are identical. Based on the prediction intervals, we find no intervention that could be expected to be unequivocally effective in improving reading scores. All interventions considered may or may not improve test scores. Some interventions show larger variability, but again it is difficult to say to what extent this reflects true variability of effects or lack of precision in the estimation of the between-study variance.

Figure 3: Meta-analyses of impacts on mathematics test scores

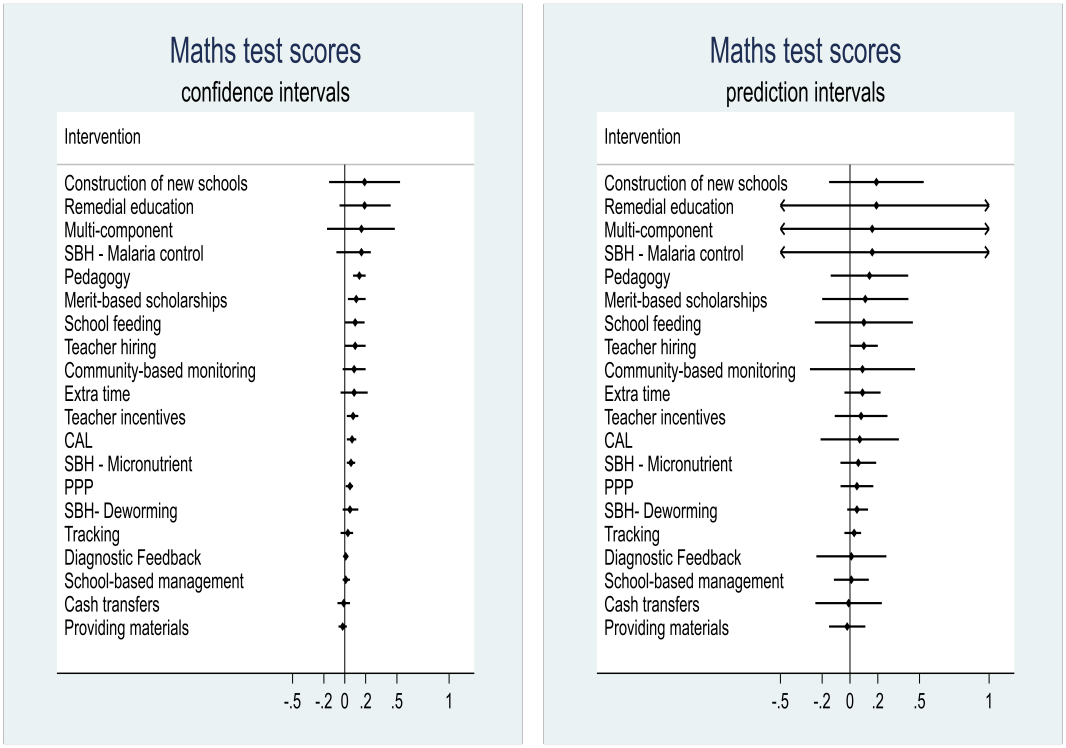


Figure 4: Meta-analyses of impacts on reading test scores

To conclude, our analysis shows that the heterogeneity of impacts of education interventions is very high for all outcomes. There is no intervention which is predictably more effective or promising than other interventions. It is also likely that our analysis underestimates the true heterogeneity for two reasons. First, for many interventions, very few studies are available and the between-study variance is imprecisely estimated. Some interventions have only two or three studies, and given the heterogeneity of effects observed for interventions with many study, we should expect the variance to increase rather than decrease after including more studies. Second, the studies included in the meta-analyses were not randomly drawn from a population of studies. Some of the studies estimated impacts of the same programme, or were conducted in the same country, or were analysed by the same researchers. A new independent study is unlikely to be very similar to the studies already included. Publication bias may also have increased the homogeneity of the studies, if journals and researchers are more likely to publish positive effects. In summary, we could expect the impact of a new study to be well off the range of the prediction interval. The heterogeneity of impacts is therefore even larger than estimated by the prediction intervals, and predictions are even more difficult to trust.

4 Conclusions

Are we able to say which education interventions are more effective in developing countries? Our analysis found that there are too many intervention types and too many differences in interventions and between the results of different studies to answer this question. The assumption of a common effect for categories of programmes is untenable and simple generalisations are not possible. Accordingly, we set out to achieve the less ambitious goal of defining, for each intervention type, the range of possible results that we could obtain by studying a new and similar intervention. The results of this exercise are not encouraging. We were not able to find and intervention more effective than others. All interventions may or may not work, and the range of possible effects for each intervention is very wide, ranging from no effects, to large effects and to negative effects. In other words, the answer to the question about what works in education in international development is that 'it depends.'

In our analysis we used the largest available dataset of studies, but this still includes relatively few studies for each intervention. Would the results of our analysis change with a larger number of studies? Much of the evidence on education interventions come from a relatively small group of interventions including conditional cash transfers, structured pedagogy and computer-assisted learning. Other interventions rely on just a handful of studies or even less. More studies of the effects of these intervention would certainly help. However, we found a high heterogeneity of results even for studies for which much evidence is available, which suggests that increasing the number of studies would not reduce the heterogeneity of the results. In fact, we have discussed how a larger number of studies

may lead to an increase in heterogeneity, as the studies already included are an artificially homogeneous group. The artificial homogeneity is a result of the research process and of publication bias, whereby studies tend to be conducted in similar conditions in the same countries, and studies producing similar (positive) effect size tend to be published.

Education interventions are very different from each other. Would a redefinition of the existing categories reduce the heterogeneity observed? Some intervention categories are very broadly defined and some interventions include various activities, in such a way that inclusion of an intervention in one category or another implies some subjective decisions. The implementation of a larger number of studies, together with a narrowing of the intervention categories to better reflect project activities, may help reducing existing heterogeneity and increasing our understanding of what works.

Finally, a last question is whether, given the heterogeneity observed, reviews of evidence and meta-analyses are a useful exercise. The answer is yes, provided quantitative reviews of evidence ask the right questions and interpret their results correctly. First, the identification of large heterogeneity is in itself a very important finding. It suggests that generalisations are not possible and that the effectiveness of the interventions cannot be understood without a deeper understanding of the context and of the characteristics of implementation. Second, much could be learned by exploring heterogeneity across studies, understanding its sources and how it relates to the results of the interventions. Indeed, the exploration of heterogeneity and its understanding should be the primary objective of a random-effects meta-analysis. On the other hand, a fixed effect meta-analysis and attempts to identify 'overall effects' of education interventions using random-effects meta-analyses are vain exercises, which can also be misleading if they falsely generate certainty about the expected impact of interventions.

References

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2009). *Introduction to Meta-Analysis*. Wiley, Chichester.

Conn, K. (2014). *Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-analysis of Rigorous Impact Evaluations*. PhD Dissertation. Columbia University, New York.

Cox, D. and Keogh, R. H. (2015). *Combination of data*. John Wiley and son, Hoboken.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*, 210:2–21.

Evans, D. and Popova, A. (2016). What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews. *World Bank Observer*, 31(2):242–270.

Glass, G. and Lee, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1(1):2–16.

Higgins, J. and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. The Cochrane Collaboration.

Higgins, J., Thompson, S., and Spiegelhalter, D. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society*, 172(Part 1):137–159.

IntHout, J., Ioannidis, J., Rovers, M., and Goeman, J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(e010247):1–6.

Krishnaratne, S., White, H., and Carpernter, E. (2013). *Quality education for all children? What works in education in developing countries*. 3ie Working Paper 20. International Initiative for Impact Evaluation, New Delhi.

Lawry, S., Samii, C., Hall, R., Leopold, A., Hornby, D., and Mtero, F. (2014). *The impact of land property rights interventions on investment and agricultural productivity in developing countries: a systematic review*. Campbell Systematic Review 2014:1.

McEwan, P. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research*, 85:354–94.

Riley, R., Higgins, J., and Deeks, J. (2011). Interpretation of random effects meta-analyses. *BMJ Research Methods and Reporting*, 342:1–6.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Qasi-Experimental Design for Generalized Causal Inference*. Wadsworth, Belmont CA.

Simpson, A. (2017). The misdirection of public policy: comparing and combining standardised effects. *Journal of Education Policy*, 32(4):450–466.

Snilstveit, B., Stevenson, J., Menon, R., Phillips, J., Gallagher, E., Geelen, M., Jobse, H., Schmidt, T., and Jimenez, M. (2016). *The impact of education programmes on learning and school participation in low and middle income countries: a systematic review summary report*. 3ie Systematic Review Summary 7. International Initiative for Impact Evaluation (3ie), London.

Snilstveit, B., Stevenson, J., Phillips, J., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M., and Eyers, J. (2015). *Interventions for improving learning outcomes and access to education in low- and middle- income countries:a systematic review*. 3ie Systematic Reviews 24. International Initiative for Impact Evaluation (3ie), London.