How Systematic Is That Systematic Review? The Case of Improving Learning Outcomes

Description

Tweet

(copy of a blog posting by David Evans on 2015/03/02 on the World Bank Development Impact blog)

Rick Davies Comment: I have highlighted interesting bits of text in red. The conclusions, also in red, are worth noting. Andâ?¦make sure you check out the great (as often) xkcd comic at the end of the posting below :-)

â??With the rapid expansion of impact evaluation evidence has come the cottage industry of the systematic review. Simply put, a systematic review is supposed to â??sum up the best available research on a specific question.â?• We found 238 reviews in 3ieâ??s database of systematic reviews of â??the effectiveness of social and economic interventions in low- and middle- income countries,â?• seeking to sum up the best evidence on topics as diverse as the effect of decentralized forest management on deforestation and the effect of microcredit on womenâ??s control over household spending.
But how definitive are these systematic reviews really? Over the past two years, we noticed that there were multiple systematic reviews on the same topic: How to improve learning outcomes for children in low and middle income countries. In fact, we found six! Of course, these reviews arenâ??t precisely the same: Some only include randomized-controlled trials (RCTs) and others include quasi-experimental studies. Some examine only how to improve learning outcomes and others include both learning and access outcomes. One only includes studies in Africa. But they all have the common core of seeking to identify what improves learning outcomes.

Here are the six studies:

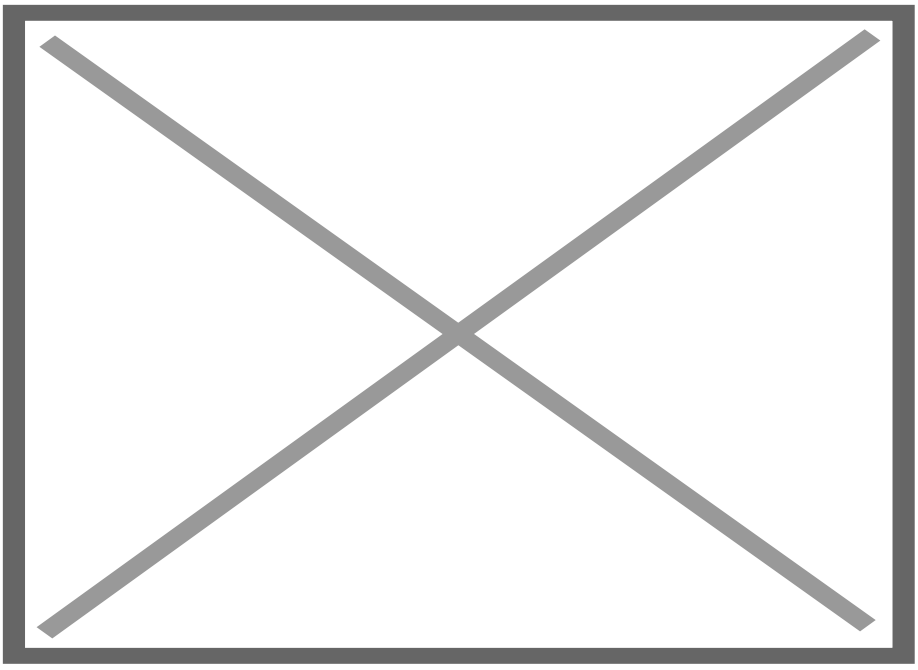
1. Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Rigorous Impact Evaluations, by Conn (2014)
2. School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990-2010, by Glewwe et al. (2014)
3. The Challenge of Education and Learning in the Developing World, by Kremer et al. (2013)
4. Quality Education for All Children? What Works in Education in Developing Countries, by Krishnaratne et al. (2013)
5. Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments, by McEwan (2014)
6. Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations, by Murnane & Ganimian (2014)

Between them, they cover an enormous amount of educational research. They identify 227 studies that measure the impact of some intervention on learning outcomes in the developing world. 134 of those are RCTs. There are studies from around the world, with many studies from China, India, Chile, and

â?? you guessed it â?? Kenya. But as we read the abstracts and intros of the reviews, there was some overlap, but also quite a bit of divergence. One highlighted that pedagogical interventions were the most effective; another that information and computer technology interventions raised test scores the most; and a third highlighted school materials as most important.

Whatâ??s going on? In a recent paper, we try to figure it out.

Differing Compositions. Despite having the same topic, these studies donâ??t study the same papers. In fact, they donâ??t even come close. Out of 227 total studies that have learning outcomes across the six reviews, only 3 studies are in all six reviews, per the figure below. That may not be surprising since there are differences in the inclusion criteria (RCTs only, Africa only, etc.). Maybe some of those studies arenâ??t the highest quality. But only 13 studies are even in the majority (4, 5, or 6) of reviews. 159 of the total studies (70 percent!) are only included in one review. 74 of those are RCTs and so are arguably of higher quality and should be included in more reviews. (Of course, there are low-quality RCTs and high-quality non-RCTs. Thatâ??s just an example.) The most comprehensive of the reviews covers less than half of the studies.



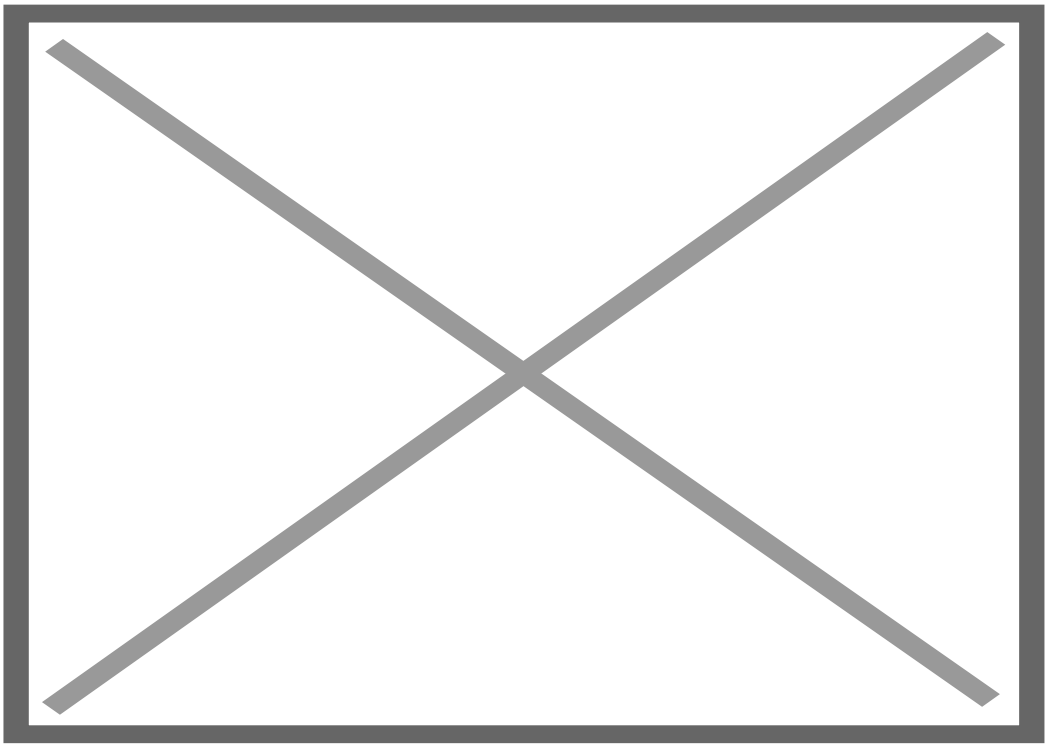
If we do a more parsimonious analysis, looking only at RCTs with learning outcomes at the primary level between 1990 and 2010 in Sub-Saharan Africa (which is basically the intersection of the inclusion criteria of the six reviews), we find 42 total studies, and the median number included in a given systematic review is 15, about one-third. So there is surprisingly little overlap in the studies that these reviews examine.

What about categorization? The reviews also vary in how they classify the same studies. For example, a program providing merit scholarships to girls in Kenya is classified alternatively as a school fee reduction, a cash transfer, a student incentive, or a performance incentive. Likewise, a program that provided computer-assisted learning in India is alternatively classified as â??computers or technologyâ?• or â??materials.â?•

What drives the different conclusions? Composition or categorization? We selected one positive recommendation from each review and examined which studies were driving that recommendation. We then counted how many of those studies were included in other reviews. As the figure below shows, the proportion varies enormously, but the median value is 33%: In other words, another review would likely have just one third of the studies driving a major recommendation in a given review. So composition matters a lot. This is why, for example, McEwan finds much bigger results for computers than others do: The other reviews include â?? on average â?? just one third of the studies that drive his result.

At the same time, categorization plays a role. One review highlights the provision of materials as one of the best ways to improve test scores. But several of the key studies that those authors call â??materials,â?• other authors categorize as â??computersâ?• or â??instructional technology.â?• While those are certainly materials, not all materials are created equal.

The variation is bigger on the inside. Systematic reviews tend to group interventions into categories (like â??incentivesâ?• or â??information provisionâ?• or â??computersâ?•), but saying that one of these delivers the highest returns on average masks the fact the variation within these groups is often as big or bigger than the variation across groups. When McEwan finds that computer interventions deliver the highest returns on average, it can be easy to forget that the same category of interventions includes a lot of clunkers, as you can see in the forest plot from his paper, below. (Weâ??re looking at you, One Laptop Per Child in Peru or in Uruguay; but not at you, program providing laptops in China. Man, thereâ??s even heterogeneity within intervention sub-categories!) Indeed, out of 11 categories of interventions in McEwanâ??s paper, 5 have a bigger standard deviation across effect sizes within the category than across effect sizes in the entire review sample. And for another 5, the standard deviation within category is more than half the standard deviation of the full sample. This is an argument for reporting effectiveness at lower levels of aggregation of intervention categories.



Source: McEwan (2014)

What does this tell us? First, itâ??s worth investing in an exhaustive search. Maybe itâ??s even worth replicating searches. Second, it may be worthwhile to combine systematic review methodologies, such as meta-analysis (which is very systematic but excludes some studies) and narrative review (which is not very systematic but allows inclusion of lots of studies, as well as examination of the specific elements of an intervention category that make it work, or not work). Third, maintain low aggregation of intervention categories so that the categories can actually be useful.

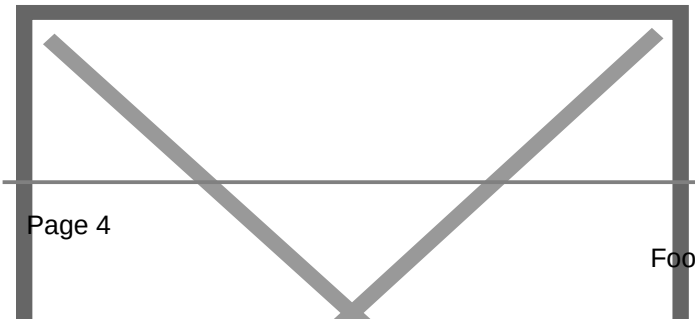
Finally, and perhaps most importantly, take systematic reviews with a grain of salt. What they recommend very likely has good evidence behind it; but it may not be the best category of intervention, since chances are, a lot of evidence didnâ??t make it into the review.

Oh, and what are the three winning studies that made it into all six systematic reviews?

1. Many Children Left Behind? Textbooks and Test Scores in Kenya, by Kremer, Glewwe, & Moulin (2009)
2. Retrospective vs. Prospective Analysis of School Inputs: The Case of Flip Charts in Kenya, by Glewwe, Kremer, Moulin, and Zitzewitz (2004)
3. Incentives to Learn, by Kremer, Miguel, & Thornton (2009)

Tomorrow, weâ??ll write briefly on what kinds of interventions are recommended most consistently across the reviews.

Future work. Can someone please now do a systematic review of our systematic review of the systematic reviews?



Credit: xkcd

Category

1. Uncategorized

Date
10/06/2026
Date Created
03/03/2015
Author
admin