American Economic Review: Papers & Proceedings 2015, 105(5): 467–470
http://dx.doi.org/10.1257/aer.p20151015

Heterogeneous Treatment Effects in Impact Evaluation[†]

By Eva Vivalt*

The past few years have seen an exponential growth in impact evaluations. These evaluations are supposed to be useful to policymakers, development practitioners, and researchers designing future studies. However, it is not yet clear to what extent we can extrapolate from past impact evaluation results or under which conditions (Deaton 2010; Pritchett and Sandefur 2013). Further, it has been shown that even a similar program, in a similar environment, can yield different results (Duflo et al. 2012; Bold et al. 2013). The different findings that have been obtained in such similar conditions point to substantial context-dependence of impact evaluation results. It is critical to understand this context-dependence in order to know what we can learn from any impact evaluation.

While the most important reason to examine generalizability is to aid interpretation and improve predictions, it could also help direct research attention to where it is most needed. If we could identify which results would be more apt to generalize, we could reallocate effort to those topics which are less well understood.

Though impact evaluations are still rapidly increasing both in number and in terms of the amount of resources devoted to them, a few thousand are already complete. We are thus at the point where we can begin to answer the question more generally.

I do this using a large, unique dataset of impact evaluation results. These data were gathered by a nonprofit research organization that I founded, AidGrade, that seeks to determine which programs work best where. To date, it has conducted 20 meta-analyses and systematic reviews of different development programs.[1] Data gathered through meta-analyses are the ideal data with which to answer the question of how much we can extrapolate from past results, as what one would want is a large database of impact evaluation results. Since data on these 20 topics were collected in the same way, we can also look across different types of programs to see if there are any more general trends.

The rest of this paper proceeds as follows. First, I briefly discuss the framework of heterogeneous treatment effects. I then describe the data in more detail and provide some illustrative results. A fuller treatment of the topic is found in Vivalt (2015).

I. Heterogeneous Treatment Effects

I define generalizability as the ability to predict results outside of the sample.[2]

I model treatment effects as depending on the context of the intervention. Suppose we have an intervention and are investigating a particular outcome, Y. In the simplest model, context can be represented as a "contextual variable," C, which interacts with the treatment such that:

$$(1) \quad Y_j = \alpha + \beta T_j + \delta C_j + \gamma T_j C_j + \varepsilon_j,$$

where j indexes individuals in the study and T indicates treatment status.

A particular impact evaluation might estimate an equation without the interaction term:

$$(2) \qquad Y_j = \alpha + \beta' T_j + \varepsilon_j.$$

*New York University, 14A Washington Mews #303, New York, NY 10003 (e-mail: eva.vivalt@nyu.edu). I thank Edward Miguel, Hunt Allcott, David Card, Ernesto Dal Bó, Bill Easterly, Vinci Chow, Willa Friedman, Xing Huang, Steven Pennings, Edson Severnini, and seminar participants at the University of California, Berkeley, Columbia University, New York University, and the World Bank for helpful comments.

[†] Go to http://dx.doi.org/10.1257/aer.p20151015 to visit the article page for additional materials and author disclosure statement.

[1] Throughout, I will refer to all 20 as meta-analyses, but some did not have enough comparable outcomes to be included in a meta-analysis and became systematic reviews.

[2] Technically, all that can be measured is "local generalizability"—the ability to predict results in a particular out of sample group. Vivalt (2015) expands on this.

β' would then capture $\beta + \gamma C$. In this situation, when the contextual variable changes, in a new context, the observed effect attributed to the treatment also changes. This appears as low generalizability.

The example described is the simplest case. One can imagine that the true state of the world has "interaction effects all the way down."

II. Data

This paper uses a database of impact evaluation results collected by AidGrade, a US nonprofit that I founded in 2012. AidGrade focuses on gathering the results of impact evaluations and analyzing the data, including through meta-analysis. Its data on impact evaluation results were collected in the course of its meta-analyses from 2012–2014.

AidGrade's meta-analyses follow the standard stages: (i) topic selection; (ii) a search for relevant papers; (iii) screening of papers; (iv) data extraction; and (v) data analysis. In addition, it pays attention to (vi) dissemination and (vii) updating of results (AidGrade 2013). These stages are described below, but the reader is referred to Vivalt (2015) for a more detailed summary and Higgins and Green (2008) for further information about meta-analyses and systematic reviews.

The interventions that were selected for meta-analysis were selected largely on the basis of there being a sufficient number of studies on that topic.

A comprehensive literature search was carried out using a mix of the search aggregators SciVerse, Google Scholar, and EBSCO/PubMed. The online databases of J-PAL, IPA, CEGA, and 3ie were also searched for completeness. Finally, the references of any existing systematic reviews or meta-analyses were collected.

Any impact evaluation which appeared to be on the intervention in question was included, barring those in developed countries. Any paper that tried to consider the counterfactual was considered an impact evaluation. Both published papers and working papers were included. Twenty topics were covered to date: conditional cash transfers; contract teachers; deworming; financial literacy training; HIV education; improved stoves; insecticide-treated bed nets; irrigation; micro health insurance; microfinance; micronutrient supplementation; mobile phone-based reminders; performance pay; rural electrification; safe water storage; scholarships; school meals; unconditional cash transfers; water treatment; and women's empowerment programs.

The subset of the data on which I am focusing for this paper is based on those papers that passed all screening stages in the meta-analyses and are publicly available online (AidGrade 2015). The search and screening criteria were very broad and, after passing the full text screening, the vast majority of papers that were later excluded were excluded merely because they had no outcome variables in common. The small overlap of outcome variables is a surprising and notable feature of the data.

When considering the variation of effect sizes within a set of papers, the definition of the set is clearly critical. If the outcome variable is defined very narrowly (e.g., "height in centimeters"), it is clear what is being measured, and one potential source of dispersion in results is removed. On the other hand, if outcomes are defined too narrowly, there may be little overlap in outcomes between papers. Therefore, multiple coding rules were used, with this paper considering narrowly-defined outcomes. The reader is referred to Vivalt (2015) for more details.

III. Results

I will restrict attention here to discussing how the treatment effects vary within intervention-outcome, using the data's original units (e.g., percentage points).

The first thing we might care about is: if we were considering the results of a particular impact evaluation, how likely is it that the point estimate for an outcome will fall within the confidence interval of another impact evaluation's estimate for the same intervention and outcome? What is the probability that the confidence intervals of the two studies will overlap? The mean will be contained in the confidence interval about 53 percent of the time; the studies' confidence intervals will overlap approximately 83 percent of the time.

Second, how far away are the results from one another? If we were to take the mean result within a particular intervention-outcome combination, what is the average difference between that and a given study's result? Putting the absolute differences in terms of percents, the average difference is 114 percent; the median across

intervention-outcomes is 48 percent. Excluding a given study from the calculation of the mean result, to focus on prediction out of sample, the absolute differences increase to an average of 311 percent and median of 52 percent.

It should be emphasized these numbers include outcomes one may not wish to consider together, such as percentage points and rate ratios. Rate ratios, as the name suggests, comprise a ratio of two rates: the rate (e.g., of the incidence of a disease) in the treatment group in the numerator and the related rate in the control group in the denominator. A change in 0.1 of a ratio, should be interpreted differently than a change in 0.1 of an outcome measured on another scale, such as in percentage points. Vivalt (2015) uses standardized data and does further analysis.

IV. Conclusions

How much impact evaluation results generalize to other settings is an important topic, and data from meta-analyses are the ideal data with which to examine the question. With data on 20 different types of interventions, all collected in the same way, we can begin to speak a bit more generally about how results tend to vary across contexts and what that implies for impact evaluation design and policy recommendations.

This paper presents some evidence on generalizability. In Vivalt (2015) I further explore the variation in results. Overall, I find a large amount of dispersion, with most papers declining to take on the tasks that would make their findings more useful, such as: specifying a model or "causal chain" through which the intervention is supposed to work; reporting results for outcome variables that other studies also consider; or providing basic information about the context of the intervention.

There are some steps that researchers can take that may improve the generalizability of their own studies. First, just as with heterogeneous selection into treatment (Chassang, Padró i Miquel, and Snowberg 2012), one solution would be to ensure one's impact evaluation varied some of the contextual variables that we might think underlie the heterogeneous treatment effects. Given that many studies are underpowered, that may not be likely; however, large organizations and governments have been supporting more impact evaluations, providing more opportunities to explicitly integrate these analyses. Efforts to coordinate across different studies, asking the same questions or looking at some of the same outcome variables, would also help. Any subgroup analyses should be prespecified so as to avoid specification searching (Casey et al. 2012). The framing of heterogeneous treatment effects could also provide positive motivation for replication projects in different contexts: different findings would not necessarily negate the earlier ones but add another level of information.

Ultimately, knowing how much results extrapolate and when is critical if we are to know how to interpret an impact evaluation's results or apply its findings. More work is needed in this vein.

REFERENCES

AidGrade. 2013. "Process map and methodology." http://www.aidgrade.org/methodology/process-map-and-methodology (accessed March 9, 2013).

AidGrade. 2015. http://www.aidgrade.org/ (accessed January 12, 2015).

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. "Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education." University of Oxford, Center for Global Development Working Paper 321.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan." *Quarterly Journal of Economics* 127 (4): 1755–1812.

Chassang, Sylvain, Gerard Padró i Miquel, and Erik Snowberg. 2012. "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments." *American Economic Review* 102 (4): 1279–1309.

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

Duo, Esther, Pascaline Dupas, and Michael Kremer. 2012. "School Governance, Teacher Incentives and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools." National Bureau of Economic Research Working Paper 17939.

Higgins, J. P. T., and Sally Green. 2008. *Cochrane Handbook for Systematic Reviews of Interventions*. West Sussex: Wiley, 2011.

Pritchett, Lant, and Justin Sandefur. 2013. "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix." Center for Global Development Working Paper 336.

Vivalt, Eva. 2015. "How Much Can We Generalize from Impact Evaluations?" Unpublished.