How Much Can We Generalize from Impact Evaluations?

Eva Vivalt*

Stanford University

May 9, 2016

Abstract

Impact evaluations aim to predict the future, but they are rooted in particular contexts and to what extent they generalize is an open and important question. We exploit a new data set of results on a wide variety of interventions and find more heterogeneity than in other literatures. This has implications for how evidence is generated and used to inform policy.

*E-mail: vivalt@stanford.edu.

1 Introduction

Recent years have seen extraordinary growth in the use of rigorous impact evaluations in the social sciences, particularly in international development. This expansion of evidence about what works is welcome. However, if this evidence is to be useful in informing policy, we must also know to what extent the results from impact evaluations generalize to new contexts. Concerns about external validity have stimulated lively theoretical debates in economics (Deaton, 2010; Pritchett and Sandefur, 2013). Further, examples of studies which had low external validity have begun to trickle in (Bold *et al.*, 2010; Allcott, 2015). There is also growing interest in extrapolating to different contexts (Dehejia et al., 2015; Gechter, 2015; Bandiera *et al.*, 2015; Kowalski, 2016), but a motivating question has still not been answered: how much do results truly vary and what does this imply for research and for policy?

This paper provides this evidence. We use a new data set of 15,024 estimates from 635 papers on 20 types of interventions in international development, gathered in the course of meta-analysis, and find that the results reported are more heterogeneous than in other fields, such as medicine.

Impact evaluation results are widely cited in reports generated for policymaking and are often shared without much information about context, study design or even standard errors. If they were, and if policymakers updated perfectly based on this information, the dispersion of studies' results would not be an issue. However, such information is not typically provided. For example, the World Development Report, the World Bank's flagship annual publication that is widely circulated among policymakers, does not typically include confidence intervals or similar information, nor is there room for a detailed description of each study.[1] Nor is this issue limited to development; at the present time of writing, the plain language two-pagers the Campbell Collaboration publishes for policymakers also provide limited contextual information and no standard errors.[2] Details about studies' implementation and other factors are frequently sparse not just in policy reports, but also in the research papers themselves: of the studies considered in this paper, 1 in 5 did not even make clear the basic detail of who implemented the program.

In order to systematically analyze heterogeneity in studies results, a comprehen-

[1] World Development Reports from 2010-2016 were checked for standard error information and only 8 cases were found out of thousands of cited papers.

[2] Based on all the reviews posted on their website, last accessed March 16, 2016.

sive and unbiased sample of studies is needed. We use those studies that were included in meta-analyses and systematic reviews by a non-profit research institute. That organization, AidGrade, seeks to systematically understand which interventions work best where. To date, AidGrade has conducted 20 meta-analyses and systematic reviews of different development programs.[3] Data gathered through meta-analyses are the ideal data to answer the question of how much we can extrapolate from past results, and since data on these 20 topics were collected in the same way, coding the same outcomes and other variables, we can look across different types of programs to see if there are any more general trends that help to explain impact evaluation results.

A further contribution of this paper is the development of benchmarks or rules of thumb that researchers or practitioners can use to gauge the relative external validity of their own work. We discuss several metrics and show typical values across a range of interventions. Other disciplines have considered generalizability more, so we draw on the literature relating to meta-analysis, which has been most well-developed in medicine, as well as the psychometric literature on generalizability theory (Higgins and Thompson, 2002; Shavelson and Webb, 2001; Briggs and Wilson, 2007).

We find that results are much more heterogeneous than in other fields. Policy decisions would seem to be improved by careful research designs that pay attention to the issue of external validity since, as it stands, the naïve prediction of the effect of a program might differ from the actual value obtained by close to 100%.

Though this paper focuses on results of impact evaluations in development, as one of the first fields within economics with enough papers on comparable topics to do this analysis, external validity is a great concern in many fields. As more rigorous impact evaluations are completed, opportunities for research relating to external validity will continue to grow.

2 Theory

We can think of two general models of the world: in the first case, there is one true effect of a particular program and all differences between studies can be attributed

[3]Throughout, we will refer to all 20 as meta-analyses, but some did not have enough comparable outcomes for meta-analysis and became systematic reviews.

simply to sampling error. In other words:

$$Y_i = \theta + \varepsilon_i \tag{1}$$

where θ is the true effect and ε_i is the error term.

In the second case, the true effect could potentially vary from context to context. Here,

$$\begin{aligned} Y_i &= \theta_i + \varepsilon_i & (2) \\ &= \bar{\theta} + \eta_i + \varepsilon_i & (3) \end{aligned}$$

where $\bar{\theta}$ is the mean true effect size, η_i is a particular study's divergence from that mean true effect size, and ε_i is the error.

These two general models correspond to the fixed effect and random-effects models for meta-analysis, where θ_i and ε_i are assumed to both be normally distributed. We do not need to impose whether a fixed effect or random-effects model better fits the data; we can simply look to the data and estimate the degree of across-study variation, τ^2, as well as the sampling variance, σ^2. $\tau^2 = 0$ indicates complete pooling; as $\tau^2 \to \infty$, we cannot pool at all.

Random-effects models are necessary if we think there are heterogeneous treatment effects and they will turn out to be more plausible given the data. We can imagine building from the random effects model to incorporate explanatory variables, generating mixed models; we will also use mixed models to explain more of the observed heterogeneity.

It should be noted that meta-analysis also allows us to improve our estimates of any one given study's treatment effect. In the random-effects model, the estimated true treatment effect $\widehat{\theta}_i$ can be shown to be equal to:

$$\widehat{\theta}_i = \frac{\frac{Y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} \tag{4}$$

where Y_i are the estimates of effect sizes from individual studies, σ_i^2 is the sampling variance, τ^2 is the component of the variation of Y_i that is not sampling variance, and

μ is the grand mean across studies.[4] As $\widehat{\theta}_i$ depends on both Y_i and μ, it can perform better than the original study's estimate Y_i at predicting the effects of a replication, especially in the presence of large sampling variance (Stein, 1955; Efron and Morris, 1975).

3 Data

This paper uses a database of impact evaluation results collected by AidGrade, a U.S. non-profit research institute founded by the author 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, 2016a).

AidGrade's meta-analyses follow the standard stages: (1) topic selection; (2) a search for relevant papers; (3) screening of papers; (4) data extraction; and (5) data analysis. In addition, it pays attention to (6) dissemination and (7) updating of results. Here, we will discuss the selection of papers (stages 1-3) and the data extraction protocol (stage 4); more detail is provided in Appendix E.

3.1 Selection of Papers

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. Five AidGrade staff members each independently made a preliminary list of interventions for examination; the lists were then combined and searches done for each topic to determine if there were likely to be enough impact evaluations for a meta-analysis. The list remaining after excluding topics with insufficient studies was voted on by the general public online and partially randomized. Appendix E provides further detail.

A comprehensive literature search was done using a mix of the search aggregators SciVerse, Google Scholar, and EBSCO/PubMed. The online databases of the Abdul Latif Jameel Poverty Action Lab (J-PAL), Innovations for Poverty Action (IPA), the Center for Effective Global Action (CEGA), and the International Initiative for Impact Evaluation (3ie) were also searched for completeness. Finally, the references

[4] Vivalt (2016b) contains a full exposition of both the random-effects and mixed model and derivation of their estimation strategies.

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.[5] Any paper that tried to consider the counterfactual of no intervention was considered an impact evaluation. Both published papers and working papers were included. The search and screening criteria were deliberately broad. The full text of the search terms and inclusion criteria for all 20 topics in this paper are available in an online appendix as detailed in Appendix A.

3.2 Data Extraction

The subset of the data on which we focus is based on those papers that passed all screening stages in the meta-analyses. Again, 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 or did not provide sufficient data for analysis (for example, not providing data that could be used to calculate the standard error of an estimate or displaying results only graphically). The small overlap of outcome variables is a surprising and notable feature of the data. Ultimately, the data we draw upon for this paper consist of 15,024 results (double-coded and then reconciled by a third researcher) across 635 papers covering the 20 types of development program listed in Table 1.[6] Only 307 of these papers overlapped in outcomes with another paper on the same intervention. The small overlap of outcome variables is a surprising and notable feature of the data and suggests researchers should coordinate more.

When considering the variation of effect sizes within a set of papers, the definition of the set is clearly critical. Two different rules were used to define outcomes: a strict rule, under which only identical outcome variables are considered alike (*e.g.* height in centimeters), and a loose rule, under which similar but distinct outcomes are grouped into clusters (*e.g.* one study may consider a subject to have anemia if

[5]High-income countries, according to the World Bank's classification system (2015).

[6]Three titles here may be misleading. "Mobile phone-based reminders" refers specifically to SMS or voice reminders for health-related outcomes. "Women's empowerment programs" required an educational component to be included in the intervention and it could not be an unrelated intervention that merely disaggregated outcomes by gender. Finally, "micronutrient supplementation" was initially too loosely defined; this was narrowed down to focus on those providing zinc to children, but the other micronutrient papers are still included in the greater data set, with a tag, and are used to examine other issues in other papers, such as publication bias.

Table 1: List of Development Programs Covered

2012	2013
Conditional cash transfers	Contract teachers
Deworming	Financial literacy training
Improved stoves	HIV education
Insecticide-treated bed nets	Irrigation
Microfinance	Micro health insurance
Safe water storage	Micronutrient supplementation
Scholarships	Mobile phone-based reminders
School meals	Performance pay
Unconditional cash transfers	Rural electrification
Water treatment	Women's empowerment programs

their hemoglobin is less than X; another may consider a subject to have anemia if their hemoglobin is less than Y). This paper uses the strict rule wherever possible.[7]

Clearly, even under the strict rule, differences between the studies may exist, however, using two different rules allows us to isolate the potential sources of variation, and other variables were coded to capture some of this variation, such as the age of those in the sample. In total, 73 variables were coded for each paper. Additional topic-specific variables were coded for some sets of papers, such as the median and mean loan size for microfinance programs. This paper focuses on the variables held in common across the different topics. These include which method was used; if randomized, whether it was randomized by cluster; whether it was blinded; where it was (village, province, country); what kind of institution carried out the implementation; characteristics of the population; and the duration of the intervention from the baseline to the midline or endline results, among others. A full set of variables and the coding manual is available online, as detailed in Appendix A. If one were to divide the studies by all these characteristics, however, the data would usually be too sparse for analysis.

Interventions were also defined separately and coders were also asked to write a short description of the details of each program. Program names were recorded so as to identify those papers on the same program. For papers which were follow-ups, the most recent results were used for each outcome.

[7]Using the loose definition preserves more data for anemia and malaria, so for these outcomes the loose definition is used.

Most analyses in this paper use the unstandardized "raw" results data reported in papers, however, the data were also standardized to be able to provide a set of results more comparable with the literature and so as not to overweight those outcomes with larger scales in some analyses. The typical way to compare results across different outcomes is to use the standardized mean difference, defined as SMD $= \frac{\mu_1 - \mu_2}{\sigma_p}$, where μ_1 is the mean outcome in the treatment group, μ_2 is the mean outcome in the control group, and σ_p is the pooled standard deviation. The Appendix describes the alternative procedures used for generating the SMD when these data were not available. The signs of the results were also adjusted so that a positive effect size always represents an improvement.

As this paper pays particular attention to the program implementer, it is worth discussing how this variable was coded in more detail. There were several types of implementers that could be coded: governments, NGOs, private sector firms, and academics. There was also a code for "other" or "unclear". The vast majority of studies were implemented by academic research teams and NGOs. This paper considers NGOs and academic research teams together because it turned out to be practically difficult to distinguish between them in the studies, especially as the passive voice was frequently used (*e.g.* "X was done" without noting who did it).

Studies tend to report results for multiple specifications. AidGrade focused on those results least likely to have been influenced by author choices: those with the fewest controls, apart from fixed effects. Where a study reported results using different methodologies, coders were instructed to collect the findings obtained under the authors' preferred methodology; where the preferred methodology was unclear, coders were advised to follow the internal preference ordering of prioritizing randomized controlled trials, followed by regression discontinuity designs and differences-in-differences, followed by matching, and to collect multiple sets of results when they were unclear on which to include. Where results were presented separately for multiple subgroups, coders were similarly advised to err on the side of caution and to collect both the aggregate results and results by subgroup except where the author appeared to be only including a subgroup because results were significant within that subgroup. For example, if an author reported results for children aged 8-15 and then also presented results for children aged 12-13, only the aggregate results would be recorded, but if the author presented results for children aged 8-9, 10-11, 12-13, and 14-15, all subgroups would be coded as well as the aggregate result when presented.

Authors only rarely reported isolated subgroups, so this was not a major issue in practice.

A note must be made about combining data. When conducting a meta-analysis, the Cochrane Handbook for Systematic Reviews of Interventions recommends collapsing the data to one observation per intervention-outcome-paper, and we do this for generating the within intervention-outcome meta-analyses (Higgins and Green, 2011). Where results had been reported for multiple subgroups (*e.g.* women and men), we aggregated them as in the Cochrane Handbook's Table 7.7.a. Where results were reported for multiple time periods (*e.g.* 6 months after the intervention and 12 months after the intervention), we used the most comparable time periods across papers.

Finally, one paper appeared to misreport results, suggesting implausibly low values and standard deviations for hemoglobin. This observation was excluded and the paper's corresponding author contacted.

3.3 Data Description

Figure 1 summarizes the distribution of studies across interventions and outcomes. Attention will typically be limited to those intervention-outcome combinations on which we have data for at least three papers.

Table 10 in Appendix C lists the interventions and outcomes and describes their results in a bit more detail, providing the distribution of significant and insignificant results. It should be emphasized that the number of negative and significant, insignificant, and positive and significant results per intervention-outcome combination only provide ambiguous evidence of the typical efficacy of a particular type of intervention. Simply tallying the numbers in each category is known as "vote counting" and can yield misleading results if, for example, some studies are underpowered.

Table 2 further summarizes the distribution of papers across interventions and highlights the fact that papers exhibit very little overlap in terms of outcomes studied. This is consistent with the story of researchers each wanting to publish one of the first papers on a topic. Vivalt (2015) finds that later papers on the same intervention-outcome combination more often remain as working papers.

Figure 1: Number of Papers on Each Unstandardized Outcome by Intervention

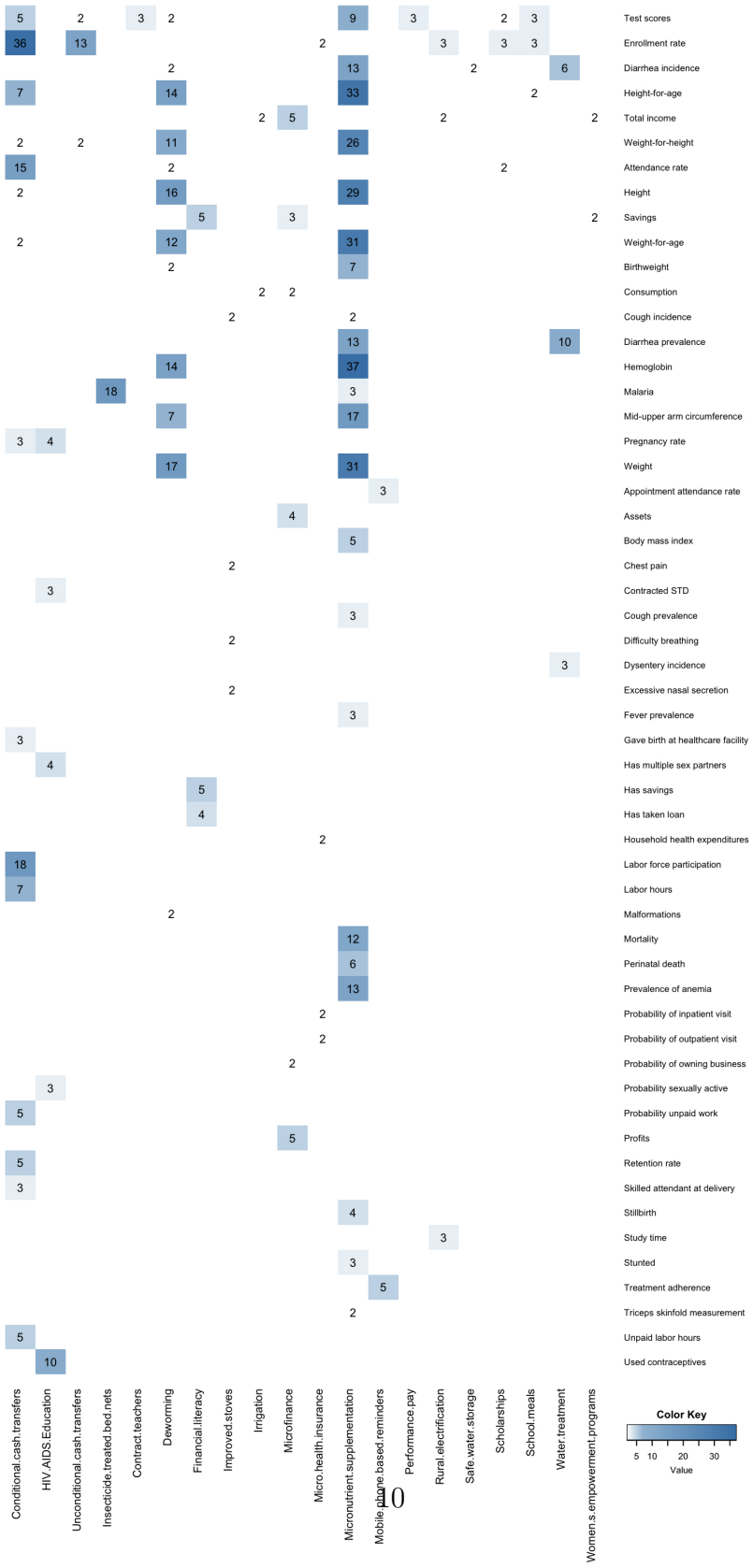

Table 2: Descriptive Statistics: Distribution of Narrow Outcomes

Intervention	Number of outcomes	Mean papers per outcome	Max papers per outcome
Conditional cash transfers	15	18	36
Contract teachers	1	3	3
Deworming	12	13	17
Financial literacy	3	5	5
HIV/AIDS Education	5	6	10
Improved stoves	4	2	2
Insecticide-treated bed nets	1	18	18
Irrigation	2	2	2
Micro health insurance	4	2	2
Microfinance	6	4	5
Micronutrient supplementation	22	23	37
Mobile phone-based reminders	2	4	5
Performance pay	1	3	3
Rural electrification	3	3	3
Safe water storage	1	2	2
Scholarships	3	2	3
School meals	3	3	3
Unconditional cash transfers	3	10	13
Water treatment	3	8	10
Women's empowerment programs	2	2	2
Average	4.8	6.6	9.1

4 Measures of Generalizability

There is a rich literature on generalizability theory, originally developed in psychometrics (*e.g.* Briggs and Wilson, 2007; Higgins and Thompson, 2002; Shavelson and Webb, 1991), but clearly also applicable to development. As there is currently no consensus in the economic literature on the best method for estimating generalizability, this paper reviews several potential measures and discusses the advantages and disadvantages of each. Acknowledging the diversity of potentially informative measures, the paper estimates the heterogeneity in each intervention-outcome combination using several different measures. Specifically, we provide measures of both the variability of results (the variance and coefficient of variation) and the proportion of

variation that can be explained (the I^2).[8]

We will also separate out the sampling variance and use explanatory variables to reduce the unexplained heterogeneity, resulting in the amount of residual variation in Y_i, the coefficient of residual variation, and the residual I^2. Appendix B has more information on these measures and motivates their use. It is important to note that each measure captures different things and has advantages and disadvantages, as summarized in Tables 8 and 9 in that section. For example, the I^2, which compares within and across-study variation, is often preferred by Bayesians, as it relates to the pooling factor (Rubin, 1981). However, it has been noted it would be artificially inflated in cases in which a study has a very large sample size and consequently small standard errors. Again, given the lack of consensus on measures, the reader is referred to Appendix B for a more complete discussion of the advantages and disadvantages of various metrics, and results for multiple measures will be presented in the text.

5 Results

5.1 Naïve Approach: Without Modeling Heterogeneity

Figures 2 and 3 summarize the variation in reported results by intervention and by intervention-outcome combination, respectively. In Figure 3, sparklines show the most recent point estimate and confidence intervals; when multiple studies were conducted for a given intervention-outcome in the same year, these were combined by random-effects meta-analysis. In general, no one intervention is clearly better at obtaining a particular outcome, and the picture provided by the most recent point estimate often changes dramatically from year to year. The confidence intervals frequently overlap, but this would be neglected if policymakers are not provided with them or do not put much weight on them.

[8]The I^2 is a measure used in the meta-analysis literature and is equal to $\frac{\tau^2}{\tau^2+\sigma^2}$. Higgins and Thompson introduce and motivate its use (2002).

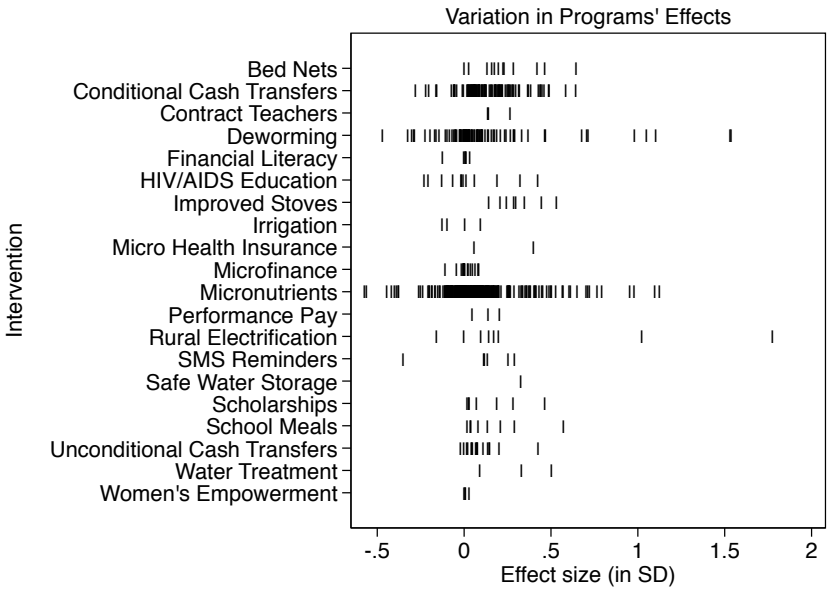

Figure 2: Dispersion of Estimates

This figure illustrates the range and density of effect sizes found for each intervention. One observation with an effect size greater than 2 is omitted for legibility.

Figure 3: Variation by Intervention-Outcome

Intervention	Outcome	Min	Max	Units	Treatment Effect
Bed Nets	Malaria	-0.77	-0.24	log risk ratio	
Conditional Cash Transfers	Attendance rate	0.02	0.10	percentage points	
Conditional Cash Transfers	Enrollment rate	0.00	0.16	percentage points	
Conditional Cash Transfers	Gave birth at healthcare facility	0.00	0.16	percentage points	
Conditional Cash Transfers	Height-for-age	-0.22	0.09	standard deviations	
Conditional Cash Transfers	Labor force participation	-0.10	0.07	percentage points	
Conditional Cash Transfers	Labor hours	-5.56	0.29	hours/week	
Conditional Cash Transfers	Pregnancy rate	-0.08	0.01	percentage points	
Conditional Cash Transfers	Probability unpaid work	-0.19	0.07	percentage points	
Conditional Cash Transfers	Retention rate	-0.04	0.00	percentage points	
Conditional Cash Transfers	Skilled attendant at delivery	0.04	0.19	percentage points	
Conditional Cash Transfers	Test scores	-0.06	0.15	standard deviations	
Conditional Cash Transfers	Unpaid labor hours	-3.94	-0.41	hours/week	
Contract Teachers	Test scores	0.14	0.27	standard deviations	
Deworming	Height	-0.64	0.84	cm	
Deworming	Height-for-age	-0.04	0.37	standard deviations	
Deworming	Hemoglobin	-0.08	0.22	g/dL	
Deworming	Mid-upper arm circumference	-0.05	0.50	cm	
Deworming	Weight	-0.28	1.30	kg	
Deworming	Weight-for-age	-0.05	0.37	standard deviations	
Deworming	Weight-for-height	-0.14	0.47	standard deviations	
Financial Literacy	Has savings	0.01	0.04	percentage points	
Financial Literacy	Has taken loan	-0.01	0.04	percentage points	
Financial Literacy	Savings	10.95	23.59	current US$	
HIV/AIDS Education	Used contraceptives	-0.07	0.13	percentage points	
Microfinance	Assets	-112.59	178.61	current US$	
Microfinance	Profits	-119.20	96.20	current US$	
Microfinance	Savings	-1.17	125.09	current US$	
Microfinance	Total income	-1.49	50.17	current US$	

Intervention	Outcome	Min	Max	Units	Treatment Effect
Micronutrients	Birthweight	0.01	0.10	kg	
Micronutrients	Body mass index	-0.03	0.35	kg/m^2	
Micronutrients	Diarrhea incidence	-0.26	0.02	log risk ratio	
Micronutrients	Diarrhea prevalence	-1.04	-0.05	log risk ratio	
Micronutrients	Height	-0.50	1.50	cm	
Micronutrients	Height-for-age	-0.04	0.38	standard deviations	
Micronutrients	Hemoglobin	-0.36	0.68	g/dL	
Micronutrients	Mid-upper arm circumference	-0.06	0.10	cm	
Micronutrients	Mortality	-1.11	0.29	log risk ratio	
Micronutrients	Perinatal death	-0.19	0.30	log risk ratio	
Micronutrients	Prevalence of anemia	-0.90	0.00	log risk ratio	
Micronutrients	Stillbirth	-0.24	0.80	log risk ratio	
Micronutrients	Stunted	-0.25	0.06	log risk ratio	
Micronutrients	Test scores	-0.10	0.23	standard deviations	
Micronutrients	Weight	-0.10	0.50	kg	
Micronutrients	Weight-for-age	-0.01	0.36	standard deviations	
Micronutrients	Weight-for-height	-0.07	0.20	standard deviations	
Performance Pay	Test scores	0.05	0.15	standard deviations	
Rural Electrification	Enrollment rate	0.07	0.09	percentage points	
Rural Electrification	Study time	-0.21	2.06	hours/day	
SMS Reminders	Appointment attendance rate	0.11	0.21	log risk ratio	
Scholarships	Enrollment rate	0.01	0.21	percentage points	
School Meals	Enrollment rate	0.02	0.26	percentage points	
School Meals	Test scores	0.02	0.29	standard deviations	
Unconditional Cash Transfers	Enrollment rate	0.01	0.19	percentage points	
Water Treatment	Diarrhea incidence	-0.92	-0.27	log rate ratio	
Water Treatment	Diarrhea prevalence	-0.77	0.02	log rate ratio	
Water Treatment	Dysentery incidence	-1.61	-0.45	log rate ratio	

Each row's black line graphically depicts the most recent estimate of Y_i for that intervention-outcome combination; the grey shaded areas, the confidence intervals. Where multiple such results are reported for a year, they are combined using random-effects meta-analysis. In years with no new data, the line is flat. The sparklines reflect what someone who took the most recent estimate as the "true" effect of the intervention would think the effect was each year. The data run horizontally from 1982 to 2014, and each row is scaled vertically for legibility with the minimum and maximum point on the black line provided. Since data are collapsed by year, this figure understates the heterogeneity. Unstandardized data are used.

Table 3: Heterogeneity Measures

	$\lvert Y_i - \widehat{Y}_i\rvert/\lvert Y_i\rvert$	CV(Y_i)	I^2
20th percentile	0.40	0.73	0.01
40th percentile	0.74	1.09	0.70
60th percentile	1.39	1.84	0.97
80th percentile	3.22	2.46	1.00*

Percentiles for each measure are calculated separately; the intervention-outcome combination at the 20th percentile for I^2, for example, need not be the intervention-outcome combination at the 20th percentile for the coefficient of variation of Y_i. The prediction error $|Y_i - \widehat{Y}_i|/|Y_i|$ is calculated by study, where $\widehat{Y}_i$ is the mean value of Y_i within that intervention-outcome combination prior to the study. The other measures are calculated by intervention-outcome combination. These are broken down by intervention-outcome in the Appendix (Table 11). The I^2 is rounded to 1 where designated with an asterisk.

Table 3 presents several key statistics. The first column provides the absolute value of the prediction error in percent terms, $|Y_i - \widehat{Y}_i|/|Y_i|$, where $\widehat{Y}_i$ is the simplest, naïve predictor of a study's result, Y_i: the mean Y_i for all studies previously completed within that intervention-outcome. The median amount by which the prediction differs from the true value is 93% using unstandardized values of Y_i or 99% using standardized values.[9] In standardized values, the average absolute value of the error is 0.18, compared to an average effect size of 0.12.

If instead of using the mean result in prior time periods as $\widehat{Y}_i$ we were to use the inverse-variance weighted meta-analysis result in prior time periods as $\widehat{Y}_i$, the median absolute percent difference between $\widehat{Y}_i$ and Y_i would be 89% using standardized data and 88% using unstandardized data, and if we were to use the median result in prior time periods, it would be 92% and 88%, respectively. These values are fairly large and there is not much difference between the results using standardized and unstandardized values.

The next set of statistics reported in Table 3 are the coefficient of variation and I^2, as estimated by hiearchical Bayesian meta-analysis. These use unstandardized values of Y_i. How should we interpret these numbers? Higgins and Thompson, who defined I^2, suggested 0.25 indicative of low, 0.5 moderate, and 0.75 high levels of heterogeneity (2002). The studies show a lot of systematic variation according to this scale. No defined benchmarks exist for the coefficient of variation, but studies in the

[9]This is the median of $|Y_i - \widehat{Y}_i|/|Y_i|$, omitting the 25 observations with $Y_i = 0$.

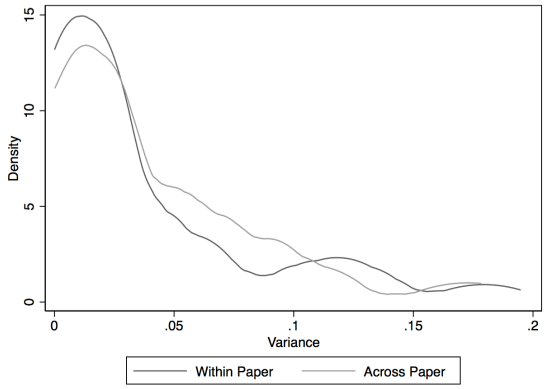

Figure 4: Distribution of Within and Across-Paper Variance

This figure plots the standardized variance of results within a paper against the variance of results across papers within an intervention-outcome combination. As there are a few outliers in the right tail, the 14 observations with variances above 0.25 are dropped for this figure.

medical literature often exhibit a coefficient of variation of approximately 0.05-0.5 (Ng, 2014; Tian, 2005). By this standard, too, results would appear quite heterogeneous. Results are broken down by intervention-outcome in Table 11.

An alternative benchmark that might have intuitive appeal is that of the average within-study variation within an intervention-outcome. If the across-study variation approached the within-study variation, we might not be so concerned about external validity across different contexts, though it should be noted that there is no guarantee within-study variation would be bounded by across-study variation. A paper might report multiple results for an intervention-outcome combination if, for example, it were reporting results for different subgroups, such as for different age groups, genders, or geographic areas. Figure 4 shows the distribution of the within-paper variance compared to the distribution of the across-paper variance. Not all studies report multiple results for an intervention-outcome combination. Within-study variation appears lower than across-study variation, but not by much. Table 12 provides additional details and compares the within-paper and across-paper coefficient of variation and I^2 measures by intervention-outcome.

Finally, we can try to derive benchmarks more directly, based on the expected prediction error. What counts as large or small error depends on the policy question. In some cases, it might not matter if a treatment effect were mispredicted by 25%.

In others, a prediction error of this magnitude could mean the difference between choosing one program over another or whether a program is worthwhile to pursue at all.

Still, if we take the mean treatment effect within an intervention-outcome to be our best guess of how a program will perform and, as an illustrative example, want the prediction error to be less than 25% at least 50% of the time, this would imply a certain cut-off threshold for the variance, assuming that results are normally distributed and the mean and variance of this distribution can be approximated by the mean and variance of the observed results.

Table 13 provides the implied bounds for $\text{var}(Y_i)$ for the prediction error to be less than 25% or 50%, respectively, at least 50% of the time, alongside the actual variance in results within each intervention-outcome. In only 2 of 57 intervention-outcome combinations is the true variance in results smaller than the variance implied by the 25% prediction error cut-off threshold, and in 14 other cases it is below the 50% prediction error threshold. In other words, for more than 70% of intervention-outcomes, the implied prediction error is greater than 50% more than 50% of the time.

While this is merely a back-of-the-envelope calculation, it highlights that in order to inform policy, modeling heterogeneity of treatment effects is of first-order importance.

5.1.1 Robustness Checks

One may be concerned that low-quality papers are either inflating or depressing the degree of heterogeneity that is observed. There are many ways to measure paper quality; two are considered here.[10]

First, we use the most widely-used quality assessment measure, the Jadad scale (Jadad *et al.*, 1996). The Jadad scale asks whether the study was randomized, double-blind, and whether there was a description of withdrawals and dropouts. A paper gets one point for having each of these characteristics; in addition, a point is added if the method of randomization was appropriate, subtracted if the method is inappropriate, and similarly added if the blinding method was appropriate and subtracted if inappropriate. This results in a 0-5 point scale. Given that the kinds of interventions

[10] Additional robustness checks can be provided on request. We acknowledge that there are many ways to measure paper quality but would argue that what is most relevant is the information provided to policymakers, and they often do not know which methods a study used, let alone receive assessments of a paper's quality.

being tested are not typically suited to blinding, we consider all those papers scoring at least a 3 to be high quality.

In an alternative specification, we also consider only those results from studies that were RCTs. This is for two reasons. First, RCTs are the gold standard in impact evaluation. Second, a separate paper finds that RCTs exhibit the fewest signs of specification searching and publication bias (Vivalt, 2015). Looking at only those studies which were RCTs thus provides a good robustness check.

Tables 14-15 provide results using the data that meet these two quality criteria. The heterogeneity measures are not substantially different using these data.

5.2 Modeling Heterogeneity

5.2.1 Across Intervention-Outcomes

If the heterogeneity in outcomes that has been observed can be systematically modeled, we might be able to make better predictions. We first look across different intervention-outcome combinations to examine whether effect sizes are associated with any characteristics of the study or sample, before turning to look within an intervention-outcome combination. To look across intervention-outcome combinations, we use standardized values.

Table 4 presents results. First, there is some evidence that studies with a smaller number of observations have greater effect sizes than studies based on a larger number of observations. This is what we would expect if specification searching were easier in small data sets; this pattern of results would also be what we would expect if power calculations drove researchers to only proceed with studies with small sample sizes if they believed the program would result in a large effect size or if larger studies are less well-targeted. Interestingly, government-implemented programs have lower effect sizes even controlling for sample size.[11] Studies in the Middle East / North Africa region may appear to perform slightly better than those in Sub-Saharan Africa (the excluded region category), but not much weight should be put on this as very few studies were conducted in the former region. RCTs do not exhibit significantly different results than quasi-experimental studies.

While looking across intervention-outcomes has the advantage of letting us draw on a larger sample of studies, and we might think that any patterns observed across so

[11]The dummy variable category left out is private sector-implemented interventions.

many interventions and outcomes would be fairly robust, we might be able to explain more variation if we restrict attention to within a particular intervention-outcome combination. We therefore focus on the case of conditional cash transfers (CCTs) and enrollment rates, as this is the intervention-outcome combination that contains the largest number of papers that should be familiar to economists.[12]

Table 4: Regression of Effect Size on Study Characteristics

	(1)	(2)	(3)	(4)	(5)
Number of observations (100,000s)	-0.013** (0.01)			-0.013** (0.01)	-0.011** (0.00)
Government-implemented		-0.081*** (0.02)			-0.073*** (0.03)
Academic/NGO-implemented		-0.018 (0.01)			-0.020 (0.01)
RCT			0.021 (0.02)		
East Asia				0.002 (0.03)	
Latin America				-0.003 (0.03)	
Middle East/North Africa				0.193** (0.08)	
South Asia				0.021 (0.04)	
Constant	0.112*** (0.00)	0.144*** (0.01)	0.093*** (0.02)	0.103*** (0.02)	0.146*** (0.01)
Observations	528	597	611	528	521
R^2	0.19	0.22	0.21	0.21	0.19

Each column reports the results of regressing the standardized effect size on different explanatory variables, dropping one outlier with an effect size greater than 2. Standard errors are clustered by intervention-outcome. Different columns contain different numbers of observations because not all studies reported each explanatory variable. Projects implemented by the private sector comprise the excluded implementer group, and the excluded region is Sub-Saharan Africa.

[12]There are more studies on micronutrients, but this is a less traditional topic for economists.

5.2.2 Within an Intervention-Outcome: The Case of CCTs and Enrollment Rates

Suppose we were to try to explain as much variation in effects of CCT programs on enrollment rates as possible using sample characteristics. The available variables which might plausibly have a relationship to effect size are: the baseline enrollment rates[13]; the sample size; whether the study was done in a rural or urban setting, or both; results for other programs in the same region[14]; and the age and gender of the sample under consideration.

Table 5 shows the results of OLS regressions of the effect size on these variables. The baseline enrollment rates show the strongest relationship to effect size, as reflected in the R^2 and significance levels: it is easier to have large gains where initial rates are low. Some papers pay particular attention to those children that were not enrolled at baseline or that were enrolled at baseline. These are coded as a "0%" or "100%" enrollment rate at baseline but are also represented by two dummy variables (Column 2). Studies done in urban areas also tend to find smaller effect sizes than studies done in rural or mixed urban/rural areas. There is no significant difference between girls and boys or based on the age of the sample.[15] Finally, for each result we calculate the mean result in the same region, excluding results from the program in question. Results do appear slightly correlated across different programs in the same region.

As baseline enrollment rates have the strongest relationship to effect size, we use this as an explanatory variable in a hierarchical mixed model (specification of Column 1), to explore how it affects the residual variance, coefficient of variation, and I^2. We also use the specification in Column 10 of Table 5 as a robustness check. The results are reported in Table 6 for each of these two mixed models, alongside the values from the random-effects model that does not use any explanatory variables.

Not all papers provide information for each explanatory variable, and each row is based on only those studies which could be used to estimate the model. Thus, the

[13]In some cases, only endline enrollment rates are reported. This variable is therefore constructed by using baseline rates for both the treatment and control group where they are available, followed by, in turn, the baseline rate for the control group; the baseline rate for the treatment group; the endline rate for the control group; the endline rate for the treatment and control group; and the endline rate for the treatment group

[14]Regions include: Latin America, Africa, the Middle East and North Africa, East Asia, and South Asia, following the World Bank's geographical divisions.

[15]Shown here: minimum sample age. Results for other age variables available upon request.

Table 5: Regression of Projects' Results on Characteristics (CCTs on Enrollment Rates)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Enrollment Rates	-0.205***	-0.102***								-0.090***
	(0.05)	(0.03)								(0.03)
Enrolled at Baseline		0.001								
		(0.02)								
Not Enrolled at		0.195***								0.199***
Baseline		(0.03)								(0.03)
Number of			-0.008							
Observations (100,000s)			(0.01)							
Rural				0.038***						0.013
				(0.01)						(0.01)
Urban					-0.049***					-0.018
					(0.01)					(0.02)
Girls						0.001				
						(0.02)				
Boys							-0.020			
							(0.01)			
Minimum Sample Age								0.001		
								(0.00)		
Mean Regional Result									1.000*	
									(0.49)	
Constant	0.212***	0.130***	0.074***	0.042***	0.071***	0.066***	0.074***	0.056***	0.000	0.115***
	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
Observations	249	249	145	270	270	270	270	244	270	249
R^2	0.32	0.44	0.00	0.05	0.03	0.00	0.01	0.00	0.02	0.45

Each column regresses the results of evaluations of conditional cash transfer programs on different explanatory variables. Multiple results for different subgroups may be reported for the same paper; the data on which this table is based includes multiple results from the same paper for different subgroups that are non-overlapping (*e.g.* boys and girls, groups with different age ranges, or different geographical areas). Standard errors are clustered by paper. Not every paper reports every explanatory variable, so different columns are based on different numbers of observations.

Table 6: Impact of Mixed Models on Measures

	$\text{var}(Y_i)$	$\text{var}_R(Y_i)$	$\text{CV}(Y_i)$	$\text{CV}_R(Y_i)$	I^2	I^2_R	N
Random effects	0.007	0.007	1.26	1.26	0.95	0.95	259
Mixed model 1	0.007	0.005	1.29	1.08	0.95	0.90	238
Mixed model 2	0.012	0.006	1.53	1.10	0.97	0.90	125

This table illustrates the impact of mixed models on heterogeneity measures. Mixed model 1 and 2 are specified by Columns 1 and 10 in Table 6. In each row, the random-effects model is used to form $\text{var}(Y_i)$, $\text{CV}(Y_i)$ and I^2; these are regenerated in each row since not all papers report the covariates used in the mixed models, so each row is based on a slightly different sample. One can see the measures of residual variation ($\text{var}_R(Y_i)$, $\text{CV}_R(Y_i)$ and I^2_R) are slightly smaller than their random-effects counterparts, although not enough to approach the medical literature.

value of $\text{var}(Y_i)$, $\text{CV}(Y_i)$ and I^2, which do not depend on the model used, may still vary between rows.

In the random-effects model, since no explanatory variables are used, $\widehat{Y}_i$ is the mean, and the measures of residual heterogeneity, $\text{var}_R(Y_i)$, $\text{CV}_R(Y_i)$ and I^2_R, do not offer improvements on $\text{var}(Y_i)$, $\text{CV}(Y_i)$ and I^2. As more explanatory variables are added, the gap between these measures grows. $\text{var}_R(Y_i)$ and $\text{CV}_R(Y_i)$ are greatly reduced from $\text{var}(Y_i)$ and $\text{CV}(Y_i)$, but I^2_R is not much lower than I^2. This is likely due to a feature of I^2 (I^2_R) previously discussed: that it depends on the precision of estimates. With evaluations of CCT programs tending to have large sample sizes, the value of I^2 (I^2_R) is higher than it otherwise would be.

This case study restricted attention to the effects of CCTs on enrollment rates. We might wonder how much better external validity would be for the other intervention-outcome combinations if we did a similar exercise for each.

It would be difficult to find great explanatory variables for each intervention-outcome combination, but we can simulate them under different assumptions. We generate the explanatory variable X_i for each result in the full data set so that regressing Y on X yields a particular R^2 within each intervention-outcome combination. Table 7 shows the results for various R^2. Based on the case of CCTs, the results for R^2=0.5 represent the preferred specification.

The main takeaway from this table is that if we had good explanatory variables, the situation would be greatly improved. We cannot compare Table 7 to the benchmarks in the medical literature, since these benchmarks were developed without modeling heterogeneity and we can imagine that the observed variation in that litera-

Table 7: Residual Heterogeneity Measures

	$\text{CV}_R(Y_i)$	I^2_R
$R^2 = 0.25$		
20th percentile	0.59	0.77
40th percentile	0.87	0.86
60th percentile	1.52	0.94
80th percentile	1.98	0.97
$R^2 = 0.50$		
20th percentile	0.51	0.70
40th percentile	0.71	0.82
60th percentile	1.18	0.92
80th percentile	1.61	0.95
$R^2 = 0.75$		
20th percentile	0.31	0.46
40th percentile	0.45	0.74
60th percentile	0.77	0.85
80th percentile	1.06	0.92

This table presents simulation results of what heterogeneity measures might look like for the entire data set if we had suitable explanatory variables to insert into a mixed model. The residual coefficient of variation, for example, is calculated using $Y_i - \widehat{Y}_i$ rather than Y_i, where $\widehat{Y}_i$ is the fitted value from the mixed model. Thus, the heterogeneity measures are capturing the residual variation that cannot be explained by the model. The mixed model is simulated by generating an explanatory variable, X_i, by taking Y_i and adding noise. The noise is normally distributed with mean 0 and a standard deviation equal to 2, 1, or 0.5 times the standard deviation of Y_i for each intervention-outcome. Each result is drawn 100 times. Percentiles for each measure are, as before, calculated separately.

ture would likewise decrease if explanatory variables were included. However, we can compare the variance to the calculated bounds for the prediction error to be less than 25% or 50%. In the preferred specification, 6 intervention-outcome combinations now have a residual variance less than the 25% prediction error cut-off threshold and 26 have one less than the 50% prediction error threshold.

6 Discussion

Why should we care about the dispersion of studies' results? If we had perfect information about the context, intervention, implementation, study quality and confidence intervals, and if we could perfectly model effect sizes and update based on that information, surely the dispersion would not matter. But that is a high bar, and it is especially unlikely a policymaker could take these factors into consideration when the reports they receive do not include this information.[16]

Reviewing World Development Reports and Campbell Collaboration "plain language" reports geared towards policymakers, the most common way of presenting information is to discursively list the point estimates of studies, sometimes noting if they are statistically significant.[17] It is up to the policymakers to determine how to weight the different studies to come up with an estimate of how they believe the program would perform in a particular setting. We must aggregate information from competing sources all the time, even in the absence of a meta-analysis or other thoughtful review.[18]

In light of this, this paper suggests that given the observed variation in studies' results, how policymakers combine information from different studies is a fruitful area for further research. Vivalt (2016a), for example, finds some evidence that people exhibit "variance neglect" in the same way they often suffer from extension neglect (sample size neglect): they do not fully take confidence intervals into consideration when updating. In this case, the point estimates form the basis of updating (approximated in this paper by using the mean point estimate to date as a predictor of the

[16]Political economy issues may also affect policymakers' decisions, but we focus on the idealized case of a policymaker who wants to make evidence-based decisions.

[17]Even then, they may not note at which level of significance. Results are also sometimes provided more qualitatively, *e.g.* "Group A showed higher levels of X than Group B", without reporting magnitudes.

[18]Thanks to Elizabeth Tipton for making this point in personal communication in 2014.

effect of the next study). There are other risks in how policymakers might view the headline results reported in policy briefs or academic papers. Without considering confidence intervals, policymakers may get the false impression that different interventions typically have different effects on particular outcomes, whereas if one were to select a random-effects meta-analysis result based on all data for an intervention-outcome and compare that to the meta-analysis result for a different intervention on the same outcome, the confidence intervals would overlap 94% of the time. If policymakers also pay more attention to the more positive results, this would lead to those interventions with a greater dispersion of results being considered to have better effects. This paper underscores the importance of further research to determine how to best present information to policymakers to enable optimal decision-making.

It is possible that variation in the precise nature of different programs classified as the "same" intervention for the purpose of analysis explains some of the observed variation in results. The reports provided to policymakers do not typically include detailed information on each intervention, so this unexplained variation is a realistic feature of the information a policymaker would receive. The confidence intervals of studies' results are also important, but again it is unlikely policymakers would perfectly update based on them if they are not even provided with them.

One might also be concerned that this paper primarily uses standardized values. As discussed, standardized values can be misleading if the standard deviation of the outcome variable differs across papers. We saw that the difference appears slight in practice, and using standardized values buys us the ability to make some additional statements looking across intervention-outcome combinations, such as in looking at the variance of Y_i and running the regression in Table 4. However, this paper's main results do not rely on using standardized values, and any result can be presented using unstandardized values upon request.

7 Conclusion

How much impact evaluation results generalize to other settings is an important question. Before now, we did not have data on many different types of interventions, all collected in the same way, with which to present a broad overview. The issues underlying external validity are well-known and assessments of external validity will always remain best conducted on a case-by-case basis. However, with the broad array

of results presented here, 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.

We consider several ways to evaluate the magnitude of the variation in results. Whether results are too heterogeneous ultimately depends on the purpose for which they are being used; some policy decisions might have greater room for error than others. However, it is safe to say that these impact evaluations exhibit more heterogeneity than is typical in other fields, such as medicine.

We also found evidence of systematic variation in effect sizes that is surprisingly robust across different interventions and outcomes. Smaller studies tended to have larger effect sizes, which we might expect if the smaller studies are better-targeted, are selected to be evaluated when there is a higher *a priori* expectation they will have a large effect size, or if there is a preference to report larger effect sizes, which smaller studies would obtain more often by chance. Government-implemented programs also had smaller effect sizes than academic/NGO-implemented programs, even after controlling for sample size. This is unfortunate given we often do smaller impact evaluations with NGOs in the hopes of finding a strong positive effect that can scale through government implementation and points to the importance of research on scaling up interventions. RCTs do not appear to have less external validity than quasi-experimental studies.

We then turn from looking across intervention-outcome combinations to explaining heterogeneity within a particular intervention-outcome combination and simulating how results using the broader data set would change if we had good explanatory variables. We find great improvement in the heterogeneity measures, underscoring that careful modeling could help substantially.

The results speak to the importance of further research on how policymakers make decisions given the information they are provided. With more research, we could find out how to improve the decisions they make.

There are also 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 as it is, 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. 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.

In summary, generalizability is not binary but something that we can measure. Policymakers should take caution when extrapolating from studies done in other contexts, and researchers should pay more attention to sampling variance, modeling, coordination, and replication.

References

AidGrade (2016a). "AidGrade Impact Evaluation Data, Version 1.3".

AidGrade (2016b). "AidGrade Process Description".

Allcott, Hunt (2015). "Site Selection Bias in Program Evaluation", Quarterly Journal of Economics.

Bandiera, Oriana, Greg Fischer, Andrea Prat and Erina Ytsma (2015). "Building Evidence from Multiple Studies: The Response to Incentive Pay", working paper.

Bold, Tessa *et al.* (2013). "Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education", working paper.

Borenstein, Michael *et al.* (2009). Introduction to Meta-Analysis. Wiley Publishers.

Briggs, Derek C., and Mark Wilson. 2007. "Generalizability in Item Response Modeling." Journal of Educational Measurement 44 (2): 13155.

Brodeur, Abel *et al.* (2012). "Star Wars: The Empirics Strike Back", working paper.

Cartwright, Nancy (2007). Hunting Causes and Using Them: Approaches in Philosophy and Economics. Cambridge: Cambridge University Press.

Cartwright, Nancy (2010). "What Are Randomized Controlled Trials Good For?", Philosophical Studies, vol. 147 (1): 59-70.

Casey, Katherine, Rachel Glennerster, and Edward Miguel (2012). "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." Quarterly Journal of Economics, vol. 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, vol. 102 (4): 1279-1309.

Cohen, Jacob (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coville, Aidan and Eva Vivalt (2016). "How Often Should We Believe Positive Results?", working paper.

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

Dehejia, Rajeev, Cristian Pop-Eleches and Cyrus Samii (2015). "From Local to Global: External Validity in a Fertility Natural Experiment", working paper.

Gechter, Michael (2015). "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India", working paper.

Gelman, Andrew *et al.* (2013). Bayesian Data Analysis, Third Edition, Chapman

and Hall/CRC.

Hedges, Larry and Therese Pigott (2004). "The Power of Statistical Tests for Moderators in Meta-Analysis", Psychological Methods, vol. 9 (4).

Higgins, Julian PT and Sally Green, *(eds.)* (2011). Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Higgins, Julian PT *et al.* (2003). "Measuring inconsistency in meta-analyses", BMJ 327: 557-60.

Higgins, Julian PT and Simon Thompson (2002). "Quantifying heterogeneity in a meta-analysis", Statistics in Medicine, vol. 21: 1539-1558.

Jadad, A.R. *et al.* (1996). "Assessing the quality of reports of randomized clinical trials: Is blinding necessary?" Controlled Clinical Trials, 17 (1): 112.

Meager, Rachel (2015). "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments", working paper.

Millennium Challenge Corporation (2009). "Key Elements of Evaluation at MCC", presentation June 9, 2009.

Ng, CK (2014). "Inference on the common coefficient of variation when populations are lognormal: A simulation-based approach", Journal of Statistics: Advances in Theory and Applications, vol. 11 (2).

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.

Pritchett, Lant, Salimah Sanji and Jeffrey Hammer (2013). "It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space", Center for Global Development Working Paper 233.

Rodrik, Dani (2009). "The New Development Economics: We Shall Experiment, but How Shall We Learn?", in What Works in Development? Thinking Big, and Thinking Small, ed. Jessica Cohen and William Easterly, 24-47. Washington, D.C.: Brookings Institution Press.

Rubin, Donald (1981). "Estimation in Parallel Randomized Experiments", Journal of Educational and Behavioral Statistics, vol. 6(4).

Shadish, William, Thomas Cook and Donald Campbell (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference.

Boston: Houghton Mifflin.
Shavelson, Richard J., and Noreen M. Webb. 1991. Generalizability Theory: A Primer. Vol. 1. Sage Publications.
Tian, Lili (2005). "Inferences on the common coefficient of variation", Statistics in Medicine, vol. 24: 2213-2220.
Vivalt, Eva (2016a). "How Do Policymakers Update?", working paper.
Vivalt, Eva (2016b). "How Much Do We Learn from an Impact Evaluation? Are They Worthwhile?", working paper.
Vivalt, Eva (2015). "The Trajectory of Specification Searching Across Disciplines and Methods", working paper.
World Bank (2015). "Country and Lending Groups", http://data.worldbank.org/about/country-and-lending-groups.
USAID (2011). "Evaluation: Learning from Experience", USAID Evaluation Policy, Washington, DC.
Young, Alwyn (2016). "Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results", working paper.

Appendices

A Guide to Appendices

A.1 Appendices in this Paper

B) Discussion of heterogeneity measures.

C) Additional results.

A.2 Further Online Appendices

Having to describe data from twenty different meta-analyses and systematic reviews, we must rely in part on online appendices. The following are available at http://www.evavivalt.com/appendices-generalize:

D) Excerpt from AidGrade's Process Description (2016b).

E) The search terms and inclusion criteria for each topic.

F) Bibliography of included and excluded papers.

G) The coding manual.

B Heterogeneity Measures

As discussed in the main text, we will want measures that speak to both the overall variability as well as the amount that can be explained.

The most obvious measure to consider is the variance of studies' results, $\text{var}(Y_i)$. A potential drawback to using the variance as a measure of generalizability is that we might be concerned that studies that have higher effect sizes or are measured in terms of units with larger scales have larger variances. This would limit us to making comparisons only between data with the same scale. We could either: 1) restrict attention to those outcomes in the same natural units (*e.g.* enrollment rates in percentage points); 2) convert results to be in terms of a common unit, such as standard deviations[19]; or 3) scale the measure, such as by the mean result, to create a unitless figure. Scaling the standard deviation of results within an intervention-outcome combination by the mean result within that intervention-outcome creates a measure known as the coefficient of variation, which represents the inverse of the signal-to-noise ratio, and as a unitless figure can be compared across intervention-outcome combinations with different natural units. It is not immune to criticism, however, particularly in that it may result in large values as the mean approaches zero.

The measures discussed so far focus on variation. However, if we could *explain* the variation, it would no longer worsen our ability to make predictions in a new setting, so long as we had all the necessary data from that setting, such as covariates, with which to extrapolate. One portion of the variation that can be immediately explained is the sampling variance, $\text{var}(Y_i|\theta_i)$, denoted σ^2. The variation in observed effect sizes is:

$$\text{var}(Y_i) = \tau^2 + \sigma^2 \tag{5}$$

and the proportion of the variation that is not sampling error is:

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{6}$$

The I^2 is an established metric in the meta-analysis literature that helps determine whether a fixed or random-effects model is more appropriate; the higher I^2, the

[19]This can be problematic if the standard deviations themselves vary but is a common approach in the meta-analysis literature in lieu of a better option.

Table 8: Summary of Heterogeneity Measures

	Measure of variation	Measure of proportion of variation that is systematic	Measure makes use of explanatory variables
$\text{var}(Y_i)$	✓		
$\text{var}_R(Y_i)$	✓		✓
$\text{CV}(Y_i)$	✓		
$\text{CV}_R(Y_i)$	✓		✓
I^2		✓	
I^2_R		✓	✓
R^2		✓	✓

less plausible it is that sampling error drives all the variation in results, and the more appropriate a random-effects model is. I^2 is considered "low" at 0.25, "moderate" at 0.5, and "high" at 0.75 (Higgins *et al.*, 2003).[20]

If we wanted to explain more of the variation, we could use a mixed model and, upon estimating it, we can calculate several additional statistics: the amount of residual variation in Y_i, after accounting for X_n, $\text{var}_R(Y_i)$, the coefficient of residual variation, $\text{CV}_R(Y_i)$, and the residual I^2_R. Further, we can examine the R^2 of the meta-regression.

It should be noted that a linear meta-regression is only one way of modeling variation in Y_i. The I^2, for example, is analogous to the reliability coefficient of classical test theory or the generalizability coefficient of generalizability theory (a branch of psychometrics), both of which estimate the proportion of variation that is not error. In this literature, additional heterogeneity is usually modeled using ANOVA rather than meta-regression. Modeling variation in treatment effects also does not have to occur only retrospectively at the conclusion of studies; we can imagine that a carefully-designed study could anticipate and estimate some of the potential sources of variation experimentally.

Table 8 summarizes the different indicators, dividing them into measures of variation and measures of the proportion of variation that is systematic.

Each of these metrics has its advantages and disadvantages. Table 9 summarizes the desirable properties of a measure of heterogeneity and which properties are pos-

[20]The Cochrane Collaboration uses a slightly different set of norms, saying 0-0.4 "might not be important", 0.3-0.6 "may represent moderate heterogeneity", 0.5-0.9 "may represent substantial heterogeneity", and 0.75-1 "considerable heterogeneity" (Higgins and Green, 2011).

Table 9: Desirable properties of a measure of heterogeneity

	Does not depend on the number of studies in a cell	Does not depend on the precision of individual estimates	Does not depend on the estimates' units	Does not depend on the mean result in the cell
$\text{var}(Y_i)$	✓	✓		✓
$\text{var}_R(Y_i)$	✓	✓		✓
$\text{CV}(Y_i)$	✓	✓	✓	
$\text{CV}_R(Y_i)$	✓	✓	✓	
I^2	✓		✓	✓
I^2_R	✓		✓	✓
R^2	✓	✓	✓	✓

A "cell" here refers to an intervention-outcome combination. The "precision" of an estimate refers to its standard error.

sessed by each of the discussed indicators. Measuring heterogeneity using the variance of Y_i requires the Y_i to have comparable units. Using the coefficient of variation requires the assumption that the mean effect size is an appropriate measure with which to scale $\text{sd}(Y_i)$. The variance and coefficient of variation also do not have anything to say about the amount of heterogeneity that can be explained. Adding explanatory variables also has its limitations. In any model, we have no way to guarantee that we are indeed capturing all the relevant factors. While I^2 has the nice property that it disaggregates sampling variance as a source of variation, estimating it depends on the weights applied to each study's results and thus, in turn, on the sample sizes of the studies. The R^2 has its own well-known caveats, such as that it can be artificially inflated by over-fitting.

To get a full picture of the extent to which results might generalize, then, multiple measures should be used.

C Additional Results

Table 10: Descriptive Statistics: Narrowly Defined Outcomes

Intervention	Outcome	# Neg sig papers	# Insig papers	# Pos sig papers	# Papers
Conditional cash transfers	Attendance rate	0	6	9	15
Conditional cash transfers	Enrollment rate	0	7	29	36
Conditional cash transfers	Gave birth at healthcare facility	0	2	1	3
Conditional cash transfers	Height	0	1	1	2
Conditional cash transfers	Height-for-age	0	6	1	7
Conditional cash transfers	Labor force participation	1	12	5	18
Conditional cash transfers	Labor hours	0	3	4	7
Conditional cash transfers	Pregnancy rate	1	1	1	3
Conditional cash transfers	Probability unpaid work	1	0	4	5
Conditional cash transfers	Retention rate	0	3	2	5
Conditional cash transfers	Skilled attendant at delivery	0	3	0	3
Conditional cash transfers	Test scores	1	2	2	5
Conditional cash transfers	Unpaid labor hours	3	2	0	5
Conditional cash transfers	Weight-for-age	0	2	0	2
Conditional cash transfers	Weight-for-height	0	1	1	2
HIV/AIDS Education	Contracted STD	0	2	1	3
HIV/AIDS Education	Has multiple sex partners	0	2	2	4
HIV/AIDS Education	Pregnancy rate	0	1	3	4
HIV/AIDS Education	Probability sexually active	0	2	1	3
HIV/AIDS Education	Used contraceptives	0	2	8	10
Unconditional cash transfers	Enrollment rate	0	5	8	13
Unconditional cash transfers	Test scores	0	1	1	2
Unconditional cash transfers	Weight-for-height	0	2	0	2
Insecticide-treated bed nets	Malaria	0	4	14	18
Contract teachers	Test scores	0	1	2	3
Deworming	Attendance rate	0	1	1	2
Deworming	Birthweight	0	2	0	2
Deworming	Diarrhea incidence	0	1	1	2
Deworming	Height	3	10	3	16
Deworming	Height-for-age	1	9	4	14
Deworming	Hemoglobin	0	13	1	14
Deworming	Malformations	0	2	0	2
Deworming	Mid-upper arm circumference	2	0	5	7
Deworming	Test scores	0	0	2	2

Deworming	Weight	3	8	6	17
Deworming	Weight-for-age	1	6	5	12
Deworming	Weight-for-height	2	7	2	11
Financial literacy	Has savings	0	4	1	5
Financial literacy	Has taken loan	0	4	0	4
Financial literacy	Savings	0	2	3	5
Improved stoves	Chest pain	0	0	2	2
Improved stoves	Cough incidence	0	0	2	2
Improved stoves	Difficulty breathing	0	0	2	2
Improved stoves	Excessive nasal secretion	0	1	1	2
Irrigation	Consumption	0	1	1	2
Irrigation	Total income	0	1	1	2
Microfinance	Assets	0	3	1	4
Microfinance	Consumption	0	2	0	2
Microfinance	Probability of owning business	0	1	1	2
Microfinance	Profits	1	3	1	5
Microfinance	Savings	0	3	0	3
Microfinance	Total income	0	3	2	5
Micro health insurance	Enrollment rate	0	1	1	2
Micro health insurance	Household health expenditures	0	1	1	2
Micro health insurance	Probability of inpatient visit	0	2	0	2
Micro health insurance	Probability of outpatient visit	0	2	0	2
Micronutrient supplementation	Birthweight	0	4	3	7
Micronutrient supplementation	Body mass index	0	1	4	5
Micronutrient supplementation	Cough incidence	0	1	1	2
Micronutrient supplementation	Cough prevalence	0	2	1	3
Micronutrient supplementation	Diarrhea incidence	0	3	10	13
Micronutrient supplementation	Diarrhea prevalence	0	5	8	13
Micronutrient supplementation	Fever prevalence	0	2	1	3
Micronutrient supplementation	Height	3	19	7	29
Micronutrient supplementation	Height-for-age	4	21	8	33
Micronutrient supplementation	Hemoglobin	6	11	20	37
Micronutrient supplementation	Malaria	0	0	3	3
Micronutrient supplementation	Mid-upper arm circumference	2	8	7	17
Micronutrient supplementation	Mortality	1	10	1	12
Micronutrient supplementation	Perinatal death	0	5	1	6

Micronutrient supplementation	Prevalence of anemia	0	0	13	13
Micronutrient supplementation	Stillbirth	0	0	4	4
Micronutrient supplementation	Stunted	0	0	3	3
Micronutrient supplementation	Test scores	1	2	6	9
Micronutrient supplementation	Triceps skinfold measurement	1	0	1	2
Micronutrient supplementation	Weight	1	17	13	31
Micronutrient supplementation	Weight-for-age	1	20	10	31
Micronutrient supplementation	Weight-for-height	0	18	8	26
Mobile phone-based reminders	Appointment attendance rate	0	0	3	3
Mobile phone-based reminders	Treatment adherence	0	2	3	5
Performance pay	Test scores	0	2	1	3
Rural electrification	Enrollment rate	0	1	2	3
Rural electrification	Study time	0	1	2	3
Rural electrification	Total income	0	2	0	2
Safe water storage	Diarrhea incidence	0	0	2	2
Scholarships	Attendance rate	0	1	1	2
Scholarships	Enrollment rate	0	2	1	3
Scholarships	Test scores	0	2	0	2
School meals	Enrollment rate	0	3	0	3
School meals	Height-for-age	0	2	0	2
School meals	Test scores	0	2	1	3
Water treatment	Diarrhea incidence	0	1	5	6
Water treatment	Diarrhea prevalence	0	3	7	10
Water treatment	Dysentery incidence	0	1	2	3
Women's empowerment programs	Savings	0	1	1	2
Women's empowerment programs	Total income	0	0	2	2
Average		0.4	3.6	3.3	7.3

Table 11: Heterogeneity Measures for Effect Sizes Within Intervention-Outcomes

Intervention	Outcome	Units	var(Y_i)	CV(Y_i)	I^2	N
Rural Electrification	Enrollment rate	percentage points	0.000	0.14	1.00	3
Financial Literacy	Has savings	percentage points	0.000	0.52	0.88	4
Conditional Cash Transfers	Retention rate	percentage points	0.000	1.12	0.98	5
Conditional Cash Transfers	Attendance rate	percentage points	0.001	0.55	1.00	14
Micronutrients	Birthweight	kg	0.002	0.99	0.96	7
Conditional Cash Transfers	Enrollment rate	percentage points	0.002	0.72	1.00	36
Conditional Cash Transfers	Labor force participation	percentage points	0.002	1.51	0.79	18
Unconditional Cash Transfers	Enrollment rate	percentage points	0.002	1.00	1.00	13
Conditional Cash Transfers	Pregnancy rate	percentage points	0.003	1.58	0.96	3
SMS Reminders	Appointment attendance rate	log risk ratio	0.003	0.32	1.00	3
Financial Literacy	Has taken loan	percentage points	0.005	2.46	0.99	4
Contract Teachers	Test scores	standard deviations	0.005	0.40	0.77	3
Conditional Cash Transfers	Skilled attendant at delivery	percentage points	0.006	0.64	0.85	3
Performance Pay	Test scores	standard deviations	0.006	0.61	1.00	3
Conditional Cash Transfers	Gave birth at healthcare facility	percentage points	0.007	1.29	0.92	3
HIV/AIDS Education	Used contraceptives	percentage points	0.008	2.22	0.14	4
Conditional Cash Transfers	Probability unpaid work	percentage points	0.009	1.52	0.92	5
Micronutrients	Weight-for-height	standard deviations	0.010	2.15	0.82	26
Micronutrients	Weight-for-age	standard deviations	0.010	1.84	0.99	31
Scholarships	Enrollment rate	percentage points	0.010	0.88	1.00	3
Micronutrients	Mid-upper arm circumference	cm	0.011	1.73	0.76	17
Micronutrients	Diarrhea incidence	log risk ratio	0.011	0.89	0.48	7
Micronutrients	Height-for-age	standard deviations	0.012	2.20	1.00	33
Conditional Cash Transfers	Test scores	standard deviations	0.013	1.87	1.00	5
School Meals	Enrollment rate	percentage points	0.016	1.14	0.97	3
Deworming	Hemoglobin	g/dL	0.018	2.21	0.46	14
Micronutrients	Body mass index	kg/mˆ2	0.022	0.68	1.00	5

Micronutrients	Test scores	standard deviations	0.023	1.93	1.00	9
School Meals	Test scores	standard deviations	0.023	1.29	0.99	3
Micronutrients	Stunted	log risk ratio	0.026	2.24	0.05	3
Micronutrients	Weight	kg	0.046	1.53	1.00	31
Conditional Cash Transfers	Height-for-age	standard deviations	0.055	22.17	0.04	7
Deworming	Weight-for-height	standard deviations	0.072	3.13	1.00	11
Deworming	Mid-upper arm circumference	cm	0.093	3.75	0.99	7
Deworming	Height-for-age	standard deviations	0.098	1.98	1.00	14
Bed Nets	Malaria	log risk ratio	0.103	0.70	0.07	10
Deworming	Weight-for-age	standard deviations	0.107	2.29	1.00	12
Micronutrients	Perinatal death	log risk ratio	0.118	2.07	0.02	6
Deworming	Weight	kg	0.153	2.66	1.00	17
Micronutrients	Prevalence of anemia	log risk ratio	0.170	0.75	0.31	13
Micronutrients	Diarrhea prevalence	log risk ratio	0.197	1.09	0.01	6
Deworming	Height	cm	0.217	5.58	0.65	16
Water Treatment	Diarrhea prevalence	log rate ratio	0.218	1.08	0.70	9
Micronutrients	Hemoglobin	g/dL	0.222	1.68	1.00	37
Micronutrients	Stillbirth	log risk ratio	0.236	3.00	0.00	4
Micronutrients	Mortality	log risk ratio	0.259	15.63	0.01	11
Water Treatment	Diarrhea incidence	log rate ratio	0.276	0.73	0.05	5
Micronutrients	Height	cm	0.309	2.88	0.84	29
Water Treatment	Dysentery incidence	log rate ratio	0.376	0.67	0.00	3
Rural Electrification	Study time	hours/day	1.382	1.06	0.01	3
Conditional Cash Transfers	Unpaid labor hours	hours/week	1.933	0.92	0.00	5
Conditional Cash Transfers	Labor hours	hours/week	4.354	1.05	0.00	7
Microfinance	Total income	current US$	512.417	0.96	0.00	5
Microfinance	Savings	current US$	5295.826	1.77	0.00	3
Microfinance	Profits	current US$	6165.543	5.45	0.00	5
Financial Literacy	Savings	current US$	9063.067	5.47	0.00	5

Microfinance	Assets	current US$	14432.350	5.51	0.00	4
Average			622.464	2.35	0.62	10
Median			0.024	1.53	0.84	6

Wherever I^2 appears equal to 1.00, this is the result of rounding.

Table 12: Across-Paper vs. Mean Within-Paper Heterogeneity

Intervention	Outcome	Across-paper var(Y_i)	Within-paper var(Y_i)	Across-paper CV(Y_i)	Within-paper CV(Y_i)	Across-paper I^2	Within-paper I^2
Conditional Cash Transfers	Attendance rate	0.001	0.001	0.55	0.53	1.00	0.96
Micronutrients	Birthweight	0.002	0.002	0.99	0.94	0.96	0.99
Conditional Cash Transfers	Enrollment rate	0.002	0.006	0.72	0.88	1.00	0.88
Conditional Cash Transfers	Labor force participation	0.002	0.003	1.51	4.30	0.79	0.96
Unconditional Cash Transfers	Enrollment rate	0.002	0.003	1.00	1.43	1.00	0.71
HIV/AIDS Education	Used contraceptives	0.008	0.394	2.22	6.97	0.14	0.07
Conditional Cash Transfers	Probability unpaid work	0.009	0.001	1.52	0.65	0.92	0.86
Micronutrients	Weight-for-height	0.010	0.005	2.15	*	0.82	0.61
Micronutrients	Weight-for-age	0.010	0.124	1.84	0.71	0.99	0.59
Scholarships	Enrollment rate	0.010	0.005	0.88	1.56	1.00	0.70
Micronutrients	Diarrhea incidence	0.011	0.064	0.89	1.88	0.48	0.51
Micronutrients	Height-for-age	0.012	0.042	2.20	3.75	1.00	0.41
Deworming	Hemoglobin	0.018	0.034	2.21	7.39	0.46	0.80
Micronutrients	Stunted	0.026	0.080	2.24	40.89	0.05	0.01
Micronutrients	Weight	0.046	0.008	1.53	0.10	1.00	0.38
Conditional Cash Transfers	Height-for-age	0.055	0.002	22.17	1.21	0.04	0.70
Deworming	Weight-for-height	0.072	0.164	3.13	*	1.00	0.97
Deworming	Height-for-age	0.098	0.005	1.98	1.84	1.00	0.80
Bed Nets	Malaria	0.103	0.209	0.70	0.61	0.07	0.50
Deworming	Weight-for-age	0.107	0.004	2.29	1.04	1.00	0.84
Micronutrients	Perinatal death	0.118	0.038	2.07	0.22	0.02	0.00
Deworming	Weight	0.153	0.116	2.66	1.89	1.00	0.57
Micronutrients	Diarrhea prevalence	0.197	0.037	1.09	6.95	0.01	0.03
Deworming	Height	0.217	0.145	5.58	0.01	0.65	0.40
Water Treatment	Diarrhea prevalence	0.218	0.334	1.08	1.01	0.70	0.08
Micronutrients	Hemoglobin	0.222	0.051	1.68	0.56	1.00	1.00
Micronutrients	Mortality	0.259	0.640	15.63	1.56	0.01	0.00
Water Treatment	Diarrhea incidence	0.276	0.117	0.73	0.56	0.05	0.19
Micronutrients	Height	0.309	0.072	2.88	3.38	0.84	0.20
Conditional Cash Transfers	Unpaid labor hours	1.933	0.951	0.92	0.85	0.00	0.00
Conditional Cash Transfers	Labor hours	4.354	5.424	1.05	3.28	0.00	0.22
Microfinance	Total income	512.417	2411.402	0.96	1.23	0.00	0.00

Within-paper values are based on those papers which report results for different subsets of the data. For closer comparison of the across and within-paper statistics, the across-paper values are based on the same data set, aggregating the within-paper results to one observation per intervention-outcome-paper. Each paper needs to have reported 3 results for an intervention-outcome combination for it to be included in the calculation, in addition to the requirement of there being 3 papers on the intervention-outcome combination. Due to the slightly different sample, the across-paper statistics diverge slightly from those reported elsewhere in the paper. Occasionally, within-paper measures of the mean equal or approach zero, making the coefficient of variation undefined or unreasonable; * denotes those coefficients of variation that were either undefined or greater than 1,000,000. Wherever I^2 appears equal to 1.00, this is the result of rounding.

Table 13: Actual Variance vs. Variance for Prediction Error Thresholds

Intervention	Outcome	Units	$\bar{Y_i}$	var(Y_i)	var_{25}	var_{50}
Rural Electrification	Enrollment rate	percentage points	0.077	0.000	0.001	0.005
Financial Literacy	Has savings	percentage points	0.024	0.000	0.000	0.001
Conditional Cash Transfers	Retention rate	percentage points	-0.012	0.000	0.000	0.000
Conditional Cash Transfers	Attendance rate	percentage points	0.069	0.001	0.001	0.004
Micronutrients	Birthweight	kg	0.042	0.002	0.000	0.002
Conditional Cash Transfers	Enrollment rate	percentage points	0.061	0.002	0.001	0.003
Conditional Cash Transfers	Labor force participation	percentage points	-0.029	0.002	0.000	0.001
Unconditional Cash Transfers	Enrollment rate	percentage points	0.050	0.002	0.000	0.002
Conditional Cash Transfers	Pregnancy rate	percentage points	-0.033	0.003	0.000	0.001
SMS Reminders	Appointment attendance rate	log risk ratio	0.179	0.003	0.005	0.028
Financial Literacy	Has taken loan	percentage points	0.028	0.005	0.000	0.001
Contract Teachers	Test scores	standard deviations	0.182	0.005	0.005	0.029
Conditional Cash Transfers	Skilled attendant at delivery	percentage points	0.123	0.006	0.002	0.013
Performance Pay	Test scores	standard deviations	0.131	0.006	0.003	0.015
Conditional Cash Transfers	Gave birth at healthcare facility	percentage points	0.065	0.007	0.001	0.004
HIV/AIDS Education	Used contraceptives	percentage points	0.041	0.008	0.000	0.001
Conditional Cash Transfers	Probability unpaid work	percentage points	-0.062	0.009	0.001	0.003
Micronutrients	Weight-for-height	standard deviations	0.045	0.010	0.000	0.002
Micronutrients	Weight-for-age	standard deviations	0.054	0.010	0.000	0.002
Scholarships	Enrollment rate	percentage points	0.114	0.010	0.002	0.011
Micronutrients	Mid-upper arm circumference	cm	0.059	0.011	0.001	0.003
Micronutrients	Diarrhea incidence	log risk ratio	-0.120	0.011	0.002	0.013
Micronutrients	Height-for-age	standard deviations	0.050	0.012	0.000	0.002
Conditional Cash Transfers	Test scores	standard deviations	0.062	0.013	0.001	0.003
School Meals	Enrollment rate	percentage points	0.111	0.016	0.002	0.011
Deworming	Hemoglobin	g/dL	0.060	0.018	0.001	0.003
Micronutrients	Body mass index	kg/mˆ2	0.218	0.022	0.007	0.041

Micronutrients	Test scores	standard deviations	0.078	0.023	0.001	0.005
School Meals	Test scores	standard deviations	0.117	0.023	0.002	0.012
Micronutrients	Stunted	log risk ratio	-0.072	0.026	0.001	0.004
Micronutrients	Weight	kg	0.140	0.046	0.003	0.017
Conditional Cash Transfers	Height-for-age	standard deviations	-0.011	0.055	0.000	0.000
Deworming	Weight-for-height	standard deviations	0.086	0.072	0.001	0.006
Deworming	Mid-upper arm circumference	cm	0.081	0.093	0.001	0.006
Deworming	Height-for-age	standard deviations	0.159	0.098	0.004	0.022
Bed Nets	Malaria	log risk ratio	-0.459	0.103	0.032	0.182
Deworming	Weight-for-age	standard deviations	0.143	0.107	0.003	0.018
Micronutrients	Perinatal death	log risk ratio	0.167	0.118	0.004	0.024
Deworming	Weight	kg	0.147	0.153	0.003	0.019
Micronutrients	Prevalence of anemia	log risk ratio	-0.549	0.170	0.046	0.260
Micronutrients	Diarrhea prevalence	log risk ratio	-0.409	0.197	0.025	0.145
Deworming	Height	cm	0.083	0.217	0.001	0.006
Water Treatment	Diarrhea prevalence	log rate ratio	-0.433	0.218	0.028	0.162
Micronutrients	Hemoglobin	g/dL	0.280	0.222	0.012	0.068
Micronutrients	Stillbirth	log risk ratio	0.162	0.236	0.004	0.023
Micronutrients	Mortality	log risk ratio	-0.033	0.259	0.000	0.001
Water Treatment	Diarrhea incidence	log rate ratio	-0.723	0.276	0.080	0.452
Micronutrients	Height	cm	0.193	0.309	0.006	0.032
Water Treatment	Dysentery incidence	log rate ratio	-0.916	0.376	0.128	0.726
Rural Electrification	Study time	hours/day	1.104	1.382	0.185	1.054
Conditional Cash Transfers	Unpaid labor hours	hours/week	-1.513	1.933	0.348	1.981
Conditional Cash Transfers	Labor hours	hours/week	-1.990	4.354	0.602	3.424
Microfinance	Total income	current US$	23.537	512.417	84.265	479.166
Microfinance	Savings	current US$	41.056	5295.826	256.374	1457.844
Microfinance	Profits	current US$	-14.414	6165.543	31.601	179.696
Financial Literacy	Savings	current US$	-17.399	9063.067	46.043	261.820

Microfinance	Assets	current US$	21.811	14432.350	72.360	411.467

var_{25} represents the variance that would result in a 25% prediction error for draws from a normal distribution centered at the mean Y_i within an intervention-outcome, $\bar{Y_i}$. var_{50} represents the variance that would result in a 50% prediction error.

Table 14: Heterogeneity Measures for RCTs

Intervention	Outcome	Units	var(Y_i)	CV(Y_i)	I^2	N
Financial Literacy	Has savings	percentage points	0.000	0.52	0.87	4
Conditional Cash Transfers	Attendance rate	percentage points	0.001	0.53	1.00	7
Micronutrients	Birthweight	kg	0.002	0.99	0.96	7
Conditional Cash Transfers	Enrollment rate	percentage points	0.002	0.63	1.00	20
Conditional Cash Transfers	Labor force participation	percentage points	0.003	1.95	0.83	7
Unconditional Cash Transfers	Enrollment rate	percentage points	0.003	0.93	1.00	8
SMS Reminders	Appointment attendance rate	log risk ratio	0.003	0.32	0.99	3
HIV/AIDS Education	Used contraceptives	percentage points	0.004	0.85	1.00	3
Financial Literacy	Has taken loan	percentage points	0.005	2.46	0.99	4
Contract Teachers	Test scores	standard deviations	0.005	0.40	0.77	3
Performance Pay	Test scores	standard deviations	0.006	0.61	1.00	3
Micronutrients	Weight-for-height	standard deviations	0.009	2.40	0.76	25
Micronutrients	Weight-for-age	standard deviations	0.010	2.07	0.98	29
Micronutrients	Diarrhea incidence	log risk ratio	0.011	0.89	0.58	7
Conditional Cash Transfers	Test scores	standard deviations	0.012	1.19	0.33	4
Micronutrients	Mid-upper arm circumference	cm	0.013	2.15	0.87	14
Micronutrients	Height-for-age	standard deviations	0.013	2.16	1.00	32
Conditional Cash Transfers	Height-for-age	standard deviations	0.013	1.44	0.11	3
Deworming	Hemoglobin	g/dL	0.018	2.21	0.43	14
Micronutrients	Test scores	standard deviations	0.021	1.45	1.00	8
School Meals	Test scores	standard deviations	0.023	1.29	1.00	3
Micronutrients	Stunted	log risk ratio	0.026	2.24	0.05	3
Micronutrients	Body mass index	kg/mˆ2	0.037	1.10	1.00	3
Micronutrients	Weight	kg	0.046	1.60	0.99	28
Deworming	Weight-for-height	standard deviations	0.072	3.13	1.00	11
Deworming	Mid-upper arm circumference	cm	0.093	3.75	1.00	7
Deworming	Height-for-age	standard deviations	0.098	1.98	1.00	14

Bed Nets	Malaria	log risk ratio	0.103	0.70	0.07	10
Deworming	Weight-for-age	standard deviations	0.107	2.29	1.00	12
Micronutrients	Perinatal death	log risk ratio	0.118	2.07	0.02	6
Deworming	Weight	kg	0.153	2.66	1.00	17
Micronutrients	Prevalence of anemia	log risk ratio	0.182	0.80	0.29	12
Micronutrients	Diarrhea prevalence	log risk ratio	0.197	1.09	0.01	6
Deworming	Height	cm	0.217	5.58	0.59	16
Water Treatment	Diarrhea prevalence	log rate ratio	0.218	1.08	0.76	9
Micronutrients	Hemoglobin	g/dL	0.228	1.72	1.00	36
Micronutrients	Stillbirth	log risk ratio	0.236	3.00	0.00	4
Micronutrients	Mortality	log risk ratio	0.259	15.63	0.01	11
Water Treatment	Diarrhea incidence	log rate ratio	0.276	0.73	0.04	5
Micronutrients	Height	cm	0.312	2.48	0.99	27
Water Treatment	Dysentery incidence	log rate ratio	0.376	0.67	0.00	3
Conditional Cash Transfers	Labor hours	hours/week	5.558	1.20	0.00	5
Financial Literacy	Savings	current US$	9063.067	5.47	0.00	5
Average			210.980	2.06	0.66	11
Median			0.026	1.45	0.87	7

Wherever I^2 appears equal to 1.00, this is the result of rounding.

Table 15: Heterogeneity Measures for Higher-Quality Studies

Intervention	Outcome	Units	var(Y_i)	CV(Y_i)	I^2	N
Financial Literacy	Has savings	percentage points	0.000	0.52	0.87	4
Micronutrients	Birthweight	kg	0.001	0.70	1.00	4
Conditional Cash Transfers	Attendance rate	percentage points	0.002	0.51	1.00	3
Unconditional Cash Transfers	Enrollment rate	percentage points	0.003	0.93	1.00	8
Conditional Cash Transfers	Enrollment rate	percentage points	0.003	0.73	1.00	10
HIV/AIDS Education	Used contraceptives	percentage points	0.004	0.85	1.00	3
Financial Literacy	Has taken loan	percentage points	0.005	2.46	0.99	4
Contract Teachers	Test scores	standard deviations	0.005	0.40	0.75	3
Conditional Cash Transfers	Test scores	standard deviations	0.006	0.58	0.88	3
Conditional Cash Transfers	Labor force participation	percentage points	0.009	2.93	0.97	3
Micronutrients	Weight-for-height	standard deviations	0.009	2.30	0.72	24
Micronutrients	Weight-for-age	standard deviations	0.010	1.85	0.99	28
Micronutrients	Diarrhea incidence	log risk ratio	0.011	0.89	0.54	7
Micronutrients	Mid-upper arm circumference	cm	0.013	1.96	0.78	14
Micronutrients	Height-for-age	standard deviations	0.013	2.05	1.00	29
Deworming	Hemoglobin	g/dL	0.018	2.21	0.49	14
Micronutrients	Test scores	standard deviations	0.021	1.45	1.00	8
School Meals	Test scores	standard deviations	0.023	1.29	1.00	3
Micronutrients	Stunted	log risk ratio	0.026	2.24	0.05	3
Micronutrients	Body mass index	kg/mˆ2	0.027	0.83	1.00	4
Micronutrients	Weight	kg	0.050	1.49	0.99	28
Deworming	Weight-for-height	standard deviations	0.072	3.13	1.00	11
Deworming	Mid-upper arm circumference	cm	0.093	3.75	0.99	7
Deworming	Height-for-age	standard deviations	0.098	1.98	1.00	14
Deworming	Weight-for-age	standard deviations	0.107	2.29	1.00	12
Bed Nets	Malaria	log risk ratio	0.135	0.89	0.23	7
Deworming	Weight	kg	0.161	2.29	1.00	15

Micronutrients	Prevalence of anemia	log risk ratio	0.170	0.75	0.28	13
Micronutrients	Perinatal death	log risk ratio	0.185	3.60	0.02	4
Micronutrients	Diarrhea prevalence	log risk ratio	0.197	1.09	0.01	6
Micronutrients	Hemoglobin	g/dL	0.220	1.66	1.00	33
Deworming	Height	cm	0.228	7.13	0.58	15
Water Treatment	Diarrhea prevalence	log rate ratio	0.232	1.02	0.77	8
Micronutrients	Mortality	log risk ratio	0.266	4.01	0.01	9
Water Treatment	Diarrhea incidence	log rate ratio	0.276	0.73	0.04	5
Micronutrients	Height	cm	0.324	2.46	0.87	26
Water Treatment	Dysentery incidence	log rate ratio	0.376	0.67	0.00	3
Conditional Cash Transfers	Labor hours	hours/week	3.848	0.58	0.00	3
Financial Literacy	Savings	current US$	9063.067	5.47	0.00	5
Average			232.572	1.86	0.69	10
Median			0.027	1.49	0.88	7

Wherever I^2 appears equal to 1.00, this is the result of rounding.