HOW MUCH CAN WE GENERALIZE FROM IMPACT EVALUATIONS?

Eva Vivalt
Australian National University

Abstract
Impact evaluations can help to inform policy decisions, but they are rooted in particular contexts and to what extent they generalize is an open question. I exploit a new data set of impact evaluation results and find a large amount of effect heterogeneity. Effect sizes vary systematically with study characteristics, with government-implemented programs having smaller effect sizes than academic or NGO-implemented programs, even controlling for sample size. I show that treatment effect heterogeneity can be appreciably reduced by taking study characteristics into account. (JEL: O21, O22, C90)

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 is welcome. However, if this evidence is to be useful in informing policy, we must also know the extent to which 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 raised questions about the external validity of initial findings have begun to trickle in (Bold et al. 2018; Allcott 2015). There is also growing interest in extrapolating to different contexts (Dehejia et al. 2019; Gechter 2015; Bandiera et al. 2016; Meager 2019; Kowalski 2016). Still, a motivating question has not yet been

The editor in charge of this paper was M. Daniele Paserman.

Acknowledgments: I am very grateful to the editor, Daniele Paserman, and three anonymous referees for useful comments and suggestions. I also thank Edward Miguel, Bill Easterly, David Card, Ernesto Dal Bó, Hunt Allcott, Suresh Naidu, Elizabeth Tipton, Vinci Chow, Willa Friedman, Xing Huang, Michaela Pagel, Steven Pennings, Edson Severnini, seminar participants at the University of California, Berkeley, Columbia University, New York University, the World Bank, Cornell University, Princeton University, the University of Toronto, the London School of Economics, the Australian National University, the University of Ottawa, the Stockholm School of Economics, Stanford University, the Inter-American Development Bank, and ASSA 2015 participants. Thanks as well to those at AidGrade, including Sampada KC, Bobbie Macdonald, Diana Stanescu, Cesar Augusto Lopez, Jennifer Ambrose, Naomi Crowther, Timothy Catlett, Joohee Kim, Gautam Bastian, Christine Shen, Taha Jalil, Risa Santoso and Catherine Razeto. Special thanks to Mi Shen for coding assistance.

E-mail: eva.vivalt@anu.edu.au

answered: how much do results truly vary and are there characteristics of studies that predict generalizability?

To answer this question, this paper leverages 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. I find a large degree of heterogeneity in treatment effects, some of which can be explained by study characteristics. In particular, smaller studies tend to report larger effect sizes, as do programs implemented by NGOs or academics. Interestingly, studies of interventions that may be thought to have a more direct causal effect seem to exhibit less heterogeneity in treatment effects, though this result is only suggestive given the small number of interventions considered. Taken together, these results suggest greater attention be paid to study and intervention characteristics. This point is worth emphasizing, since impact evaluation results are widely cited in reports generated for policymaking but are often shared without much information about context, study design or even standard errors. If researchers knew all the factors that could be affecting results and could fully explain heterogeneity in treatment effects, and if this information were included in policy reports, the dispersion of studies' results would not be an issue. However, even much more basic information is typically not provided. For example, there is not room in the World Development Report, the World Bank's flagship annual publication that is widely circulated among policymakers, for a detailed description of each study, nor do these reports typically include confidence intervals or similar information.[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 what type of organization (government, non-profit, private sector, researcher or other) implemented the program, and in more than 1 in 4 papers it was not clear how much time had elapsed between the beginning of the intervention and the collection of midline or endline data.

In order to systematically analyze heterogeneity in studies' results, a comprehensive and unbiased sample of studies is needed. I use those studies that were included in meta-analyses and systematic reviews by a non-profit research institute, AidGrade. To date, AidGrade has conducted 20 meta-analyses and systematic reviews of different development programs.[3] These meta-analyses draw upon the results reported in the initial studies. To more thoroughly model heterogeneity in treatment effects, ideally one would want micro-data from large-scale, coordinated studies covering the same outcome variables and with the same covariates collected across many different settings. However, given that micro-data are rarely available, the results data reported

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.

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

in academic papers represent the typical best option. Since the results reported in academic papers on these 20 topics were extracted from their source papers in the same way, coding the same outcomes and other variables, I can look across different types of programs to see if there are any more general trends that help to explain impact evaluation results.

Before I can begin to discuss heterogeneity in treatment effects, an introduction to Bayesian hierarchical models is warranted, as they are still quite new in economics with notable exceptions (Bandiera et al. 2016; Meager 2019). Other disciplines such as medicine and psychometrics have more thoroughly considered generalizability (e.g., Shavelson and Webb 1991; Higgins and Thompson 2002; Briggs and Wilson 2007), but there is as yet no widespread agreement on measures of heterogeneity in economics. I discuss the strengths and limitations of candidate measures of heterogeneity and explicitly tie them to generalizability using the framework of Bayesian meta-analysis. I demonstrate how these measures can help to address several key policy questions, such as: 1) given a set of results on the effect of a particular intervention (e.g., conditional cash transfers) on a particular outcome (e.g., school enrollment rates), what is the likelihood that we would accurately predict the sign of the true effect of a similar study in another context?; 2) how well can we predict the magnitude of that true effect? These questions are a simple extension of Type S and Type M errors discussed by Gelman and Carlin (2014) Gelman and Tuerlinckx (2000). Type S errors are the probability of a significant result having the incorrect sign, and Type M errors represent the magnitude by which a significant point estimate differs from the true value it seeks to estimate. While Gelman and Carlin consider replications, essentially capturing the generalizability of a study's results to its own setting, a similar approach can be leveraged to consider generalizability to another setting. I find that without considering study or intervention characteristics, an inference about another study will have the correct sign about 61% of the time for the median intervention-outcome pair in my sample. If trying to predict the treatment effect of a similar study using only the mean treatment effect in an intervention-outcome combination, the median ratio of the $\sqrt{MSE}$ to that mean is 2.49 across intervention-outcome combinations. Further, only about 6% of the observed variation in study results can be attributed to sampling variance. I find about 20% of the remaining variance could be explained using a single best-fitting explanatory variable. However, this statistic obscures a lot of heterogeneity, with the median decrease being about 10% among the intervention-outcomes for which this comparison was made. The results underscore both the large amount of true inter-study variance and the importance of careful modeling of treatment effects using micro-data.

2. Theory

Consider a set of studies on the effects of similar interventions performed in different locations or contexts; for example, studies on the effect of conditional cash transfer programs on school enrollment rates. Given a set of such studies, one may wish to

predict the true effect of the intervention in another context. I will argue that one can estimate how well one can extrapolate from a set of results using some basic measures of heterogeneity.

However, as generalizability and models from the meta-analysis literature are relatively under-considered within economics, an introduction to these models is warranted. This section will therefore be structured as follows. First, I will introduce notation and the basic models used in the meta-analysis literature. This will be followed by a discussion of how these models can be estimated. I will then introduce a set of potential heterogeneity measures relating to the model and motivate use of one particular measure, τ^2. I will first motivate its use by considering how it has been used in the literature to improve estimates of a study's true effect in that study's own setting. Finally, I will show that the same approach can be used to make inferences about the true effect of similar studies in other settings.

2.1. Bayesian Meta-Analysis

The meta-analysis literature suggests two general types of models that can be parameterized in many ways: fixed-effect models and random-effects models.[4]

Fixed-effect models assume there is one true effect of a particular program and all differences between studies can be attributed simply to sampling error. In other words:

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

where Y_i is the point estimate in study i, θ is the true effect and ε_i is the error term.

Random-effects models do not make this assumption; the true effect could potentially vary from context to context. Here,

$$Y_i = \theta_i + \varepsilon_i \tag{2}$$

where θ_i is the true effect. Random-effects models are more suitable than fixed-effect models when there are heterogeneous treatment effects and they are also more plausible. Random-effects models can also be modified by the addition of explanatory variables, at which point they are called mixed models. Both random-effects models and mixed models will be considered in this paper, however, to build intuition I will focus the exposition on the random-effects case.

A common approach taken to estimate the random-effects model is to weight each study's point estimate by the inverse of the variance of the estimate, using the standard error associated with the estimate. In my analyses, I will instead take a fully Bayesian approach. In particular, I will assume:

$$\theta_i \sim N(\mu, \tau^2) \tag{3}$$

$$Y_i|\theta_i \sim N(\theta_i, \sigma_i^2) \tag{4}$$

4. Much of this exposition will draw from Gelman et al. (2013), and the interested reader is also referred to Borenstein et al. (2009) for a gentle introduction to meta-analysis.

where μ and τ^2 are unknown hyperparameters and σ_i^2 is the sampling variance, assumed known. There are two sources of variation in this model: the true inter-study variation, τ^2, and the sampling variance, σ_i^2. Equation 4 is generally justified by considering that in a given study, sample sizes are large and so the central limit theorem holds. For a large enough study, one might be confident in assuming the sampling variance known, though in principle it could be estimated with some noise. The top level in the hierarchy, represented in Equation 3, is more controversial. θ_i can alternatively be assumed to follow other distributions, and a nonparametric approach could even be taken. However, I would like to pick one workhorse model that can be broadly applied across many intervention-outcome combinations, as that may help in interpreting the variance term, and a normal distribution seems best for this purpose. I will later perform posterior predictive checks to gauge the suitability of this model for each of the intervention-outcome combinations I study.

In practice, a researcher will observe Y_i, the study's point estimate, and its standard error. Y_i and σ_i^2 are taken as known, with the standard error conventionally used to estimate σ_i.[5] The other parameters, θ_i, μ and τ^2, will have to be estimated. There is a large literature on estimating these models (e.g., Gelman 2006; Rubin 1981; Efron and Morris 1975). I outline a simple approach to estimating a fully Bayesian model, following Gelman et al. (2013).

2.2. Estimating a Bayesian Hierarchical Random-Effects Model

Bayes' rule says that the posterior probability is proportional to the likelihood of the data given certain parameter values multiplied by the prior probability of those parameters. Ultimately, I will want to estimate the parameters θ (a vector of θ_i), μ and τ given the data. I do this by making draws from the joint posterior distribution $p(\theta, \mu, \tau|Y)$. Note that $p(\theta, \mu, \tau|Y)$ can be written as $p(\theta|\mu, \tau, Y)p(\mu|\tau, Y)p(\tau|Y)$. In estimating the model, I will draw the hyperparameters τ, then μ, from their marginal posterior distributions and draw θ from its posterior distribution conditional on the drawn values of μ and τ. The rest of this section follows Gelman et al. (2013) in writing down the posterior distributions for $p(\theta|\mu, \tau, Y)$, $p(\mu|\tau, Y)$, and $p(\tau|Y)$ that will be used.

If there are N studies i in a given intervention-outcome combination, $p(\theta|\mu, \tau, Y)$ factorizes into N components:

$$p(\theta|\mu, \tau, Y) = \prod_i p(\theta_i|\mu, \tau, Y) \tag{5}$$

Equation 3 provides the prior for θ_i, where μ and τ are unknown hyperparameters that will need to be estimated and Equation 4 provides the likelihood. Conditioning on

5. It should be noted that the standard error may estimate σ_i only with some noise. I will not be able to assess this in my data, but the approximation is generally considered unlikely to be problematic (Gelman et al. 2013) and if the fit were really poor this would show up in the fit of the model, which I will check using posterior predictive checks.

the distribution of the data, the posterior is:

$$\theta_i|\mu,\tau,Y \sim N(\widehat{\theta_i}, V_i) \tag{6}$$

where

$$\widehat{\theta_i} = \frac{\frac{Y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}, V_i = \frac{1}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} \tag{7}$$

and Y is a vector of all Y_i within the intervention-outcome combination.

I will assume a uniform prior for $\mu|\tau$ following Gelman et al. (2013) and update based on the data. As the Y_i are estimates of μ with variance $(\sigma_i^2 + \tau^2)$, the posterior of μ is given by:

$$\mu|\tau,Y \sim N(\widehat{\mu}, V_\mu) \tag{8}$$

where

$$\widehat{\mu} = \frac{\sum_i \frac{Y_i}{\sigma_i^2+\tau^2}}{\sum_i \frac{1}{\sigma_i^2+\tau^2}}, V_\mu = \frac{1}{\sum_i \frac{1}{\sigma_i^2+\tau^2}} \tag{9}$$

For τ, I again use a uniform prior over a large range of possible values. To obtain $p(\tau|Y)$, first note that $p(\tau|Y) = p(\mu,\tau|Y)/p(\mu|\tau,Y)$. The denominator of this equation is given by Equation 8; for the numerator, $p(\mu,\tau|Y)$ is proportional to $p(\mu,\tau)p(Y|\mu,\tau)$ and the marginal distribution of $Y_i|\mu,\tau$ is known:

$$Y_i|\mu,\tau \sim N(\mu, \sigma_i^2 + \tau^2) \tag{10}$$

Hence, for the numerator:

$$p(\mu,\tau|Y) \propto p(\mu,\tau)\prod_i N(Y_i|\mu,\sigma_i^2+\tau^2) \tag{11}$$

Substituting into the equation for $p(\tau|Y)$, this yields:

$$p(\tau|Y) \propto \frac{p(\tau)\prod_i N(Y_i|\mu,\sigma_i^2+\tau^2)}{N(\mu|\widehat{\mu},V_\mu)} \tag{12}$$

As this equation must hold for any μ, including $\widehat{\mu}$, $\widehat{\mu}$ can be substituted for μ, and it is this expression that I will evaluate:

$$p(\tau|Y) \propto \frac{p(\tau)\prod_i N(Y_i|\widehat{\mu},\sigma_i^2+\tau^2)}{N(\widehat{\mu}|\widehat{\mu},V_\mu)} \tag{13}$$

$$\propto p(\tau)V_\mu^{1/2}\prod_i(\sigma_i^2+\tau^2)^{-1/2}\exp\left(-\frac{(Y-\widehat{\mu})^2}{2(\sigma_i^2+\tau^2)}\right) \tag{14}$$

Given the equations for the posteriors, estimating the parameters is merely a matter of making simulations. First, I approximate a uniform distribution for the prior of τ

by generating 2,000 equally spaced points over a large range.[6] Then, I sample from the posterior of $\tau|Y$, $\mu|\tau, Y$ and $\theta_i|\mu, \tau, Y$, 10,000 times. R code implementing this approach is included in an appendix.

It should be noted that all these calculations are done within each intervention-outcome combination, independently. It would be possible to analyze data at the intervention level instead or add a level to the model such that the mean true effect of one intervention-outcome combination can be informative about the mean true effect of another intervention-outcome combination. Pooling results across outcomes within an intervention would have the benefit of increasing the number of observations that could be leveraged in an analysis; it will be shown that not many studies cover the same intervention-outcome combination. It could also mitigate the issue that multiple results on different outcomes are sometimes taken from the same paper, leading to dependence between intervention-outcome combinations. However, this approach also has a major drawback: my focus in this paper is on estimating heterogeneity, and these estimates could be artificially inflated by pooling across diverse outcomes. Further, I will be wanting to explain the observed variation in treatment effects, and some of the explanatory variables may have different relationships with different outcomes. Aggregating the data could make it harder to explain the variation. It could also make the source of the results less transparent. I will thus present results separately for each intervention-outcome combination as a conservative approach, but caution should be taken in interpreting results across intervention-outcomes given that they are correlated.

2.3. Estimating a Mixed Model

One way to extend the basic random-effects model would be by adding explanatory variables, making it a mixed model. The estimation strategy is similar. Here, as the simplest model, I will assume:

$$Y_i = \alpha + X_i\beta + e_i + u_i \tag{15}$$

with $e_i \sim N(0, \tau^2)$ capturing the true unexplained variance between studies and $u_i \sim N(0, \sigma_i^2)$ capturing the sampling error. Again, posterior distributions will be constructed from the priors and likelihood functions for each parameter to be estimated. Online Appendix D contains a derivation of the relevant posterior distributions, which are similar to the posterior distributions used in the random-effects model, and the estimation procedure is analogous. To estimate the parameters, I will again start by generating the uniformly-distributed prior for τ over a large range, then sampling from the posterior of $\tau|Y$, $\beta|\tau, Y$ and $e_i|\beta, \tau, Y$. It should be noted that the τ^2 that is estimated here will be smaller than the τ^2 estimated using a random-effects model, since some of the variance in Y_i will have been explained by X_i. For clarity, I will

6. Specifically, for each intervention-outcome I generate the standard deviation of point estimates and generate 2,000 points spaced equally between 0 and 10*sd.

henceforth denote this "residual" τ^2 as τ^2_R. As with the random-effects model, sample R code for implementing the estimation procedure is included in an appendix.

2.4. Heterogeneity Measures

As a measure of the true inter-study variation, τ^2 may be an attractive measure of heterogeneity. But there are many qualities one might like a measure of heterogeneity to have. This section discusses desirable properties for a measure of heterogeneity and shows how τ^2 compares to other potential measures that could be used.

First, it should be noted that some measures capture the variability of results and some measure the proportion of variation that can be explained. Both types of measures can be important: if the variation can be explained, it may not be a problem in making inferences; on the other hand, if overall variation is large, then even explaining a large proportion of it may result in inaccurate predictions.

Within the first category, the most obvious measure to consider is the variance of studies' results within a given intervention-outcome combination, var(Y_i). A potential drawback to using this measure is that studies that have larger effects or are measured in terms of units with larger scales will have larger variances. One can only make comparisons between data with the same scale. Hence, the literature suggests either: (1) restricting attention to those outcomes that have the same natural units (e.g., enrollment rates in percentage points); (2) converting results to be in terms of a common unit, such as standard deviations; or (3) scaling the measure, such as by the mean result, to create a unitless figure. Each approach has drawbacks. Restricting attention to outcomes in the same natural units can be limiting. Converting results to be in terms of standard deviations can be problematic if the standard deviations themselves vary, but it is a common approach in the meta-analysis literature. Scaling the standard deviation of results within an intervention-outcome combination by the mean result within that intervention-outcome creates a unitless measure known as the coefficient of variation (CV), which represents the inverse of the signal-to-noise ratio. As a unitless figure, this measure can be compared across intervention-outcome combinations with different natural units, however, it is not immune to criticism, particularly in that it may result in large values as the mean approaches zero.

The measures discussed so far focus on variation. However, if the variation could be *explained*, it would no longer result in inaccurate predictions in a new setting. As mentioned, the variation in observed treatment effects is:

$$\text{var}(Y_i) = \tau^2 + \sigma_i^2 \tag{16}$$

where τ^2 represents the true inter-study variation and σ_i^2 is the sampling variance. τ^2 thus represents the maximum inter-study variation that could be explained by a model. As the true inter-study variation, τ^2 could be an attractive measure of heterogeneity, however, it suffers from the same problem as var(Y_i) in that it depends on the outcome's units.

One may also be interested in the proportion of the variation that is not sampling

TABLE 1. Desirable properties of a measure of heterogeneity.

	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
$\text{var}(Y_i)$	✓		✓
$\text{CV}(Y_i)$	✓	✓	
τ^2	✓		✓
I^2		✓	✓

Notes: The "precision" of an estimate refers to its standard error. A "cell" here refers to an intervention-outcome combination. Each measure could be applied to summarize the remaining variation after fitting the data to a more complicated model.

error. A common such measure is:

$$I^2 = \frac{\tau^2}{\tau^2 + s^2} \tag{17}$$

where s^2 is a measure of the sampling variance that is taken to be held in common across a set of studies.[7]

The I^2 statistic is an established unitless metric in the meta-analysis literature that helps determine whether a fixed or random-effects model is more appropriate. The higher the I^2, the less plausible it is that sampling error drives all the variation in results, and the more appropriate a random-effects model is. While I^2 has the nice property that it is unitless and 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. To get a full picture of the extent to which results might generalize, then, multiple measures may be helpful.

In short, each of these statistics has its advantages and disadvantages. Table 1 summarizes which of the desirable properties of a measure of heterogeneity are possessed by each of the proposed measures. Of these measures, a Bayesian may prefer measures that separate out sampling variance, such as τ^2 or I^2. While I^2 depends on the sampling variance, a Bayesian might consider this an advantage rather than a disadvantage, as it tells us something about the informativeness of a result. To further motivate the focus on τ^2, and to a lesser extent I^2, I will describe a couple of situations in which these measures may be particularly useful.

7. Higgins and Thompson (2002) in their seminal paper defining the I^2 statistic, take a weighted mean of the

$$s^2 = \frac{(k-1)\sum \frac{1}{\sigma_i^2}}{\left(\sum \frac{1}{\sigma_i^2}\right)^2 - \sum \left(\frac{1}{\sigma_i^2}\right)^2},$$

where k is the number of studies. When there is a small number of studies, this may serve to slightly depress s but represents the conventional approach to estimating I^2.

2.5. Leveraging Heterogeneity Measures to Improve Estimates

First, consider a lab experiment conducted in several settings, as in the "Many Labs" project (Klein et al. 2014). Each experiment has high "internal validity", defined as the ability to identify the causal effect of the treatment (Banerjee and Duflo 2009). One can also look across experiments to gauge the "external validity" of one set of n results to another setting, i.e., how well observed point estimates $Y_1, Y_2, \ldots, Y_n$ can be used to jointly predict the point estimate of another study, Y_j, or the true underlying effect θ_j, perhaps in conjunction with a more complicated model.

Importantly, the best estimate of θ_j is not Y_j. Rather, Y_j may be improved upon by considering information external to study j, i.e., data from other studies $i = 1, \ldots, n$. For example, it is possible that study j, while unbiased, had a very small sample size. To the extent to which the other studies are informative about the effect in this setting, one would want to leverage those data to improve the estimate of θ_j. The $\widehat{\theta}_j$ that is estimated from the Bayesian model is a "shrinkage estimator", and the degree of shrinkage depends on the precision of the estimate relative to τ^2:

$$\widehat{\theta}_j = (1 - \lambda_j)Y_j + \lambda_j \mu \tag{18}$$

where $\lambda_j = \sigma_j^2/(\sigma_j^2 + \tau^2)$. These estimators have a storied past (e.g., Rubin 1981; Efron and Morris 1975; Stein 1956).

In this example, knowing the relationship between sampling variance and τ^2 is clearly helpful and can improve estimates of what a replication would find in the same setting. In effect, this example can be thought of as giving us the generalizability of a result to its own setting.

As a second motivating example, consider the case in which one would like to know whether two parameters, θ_j and θ_k, are equal. This could be thought of as testing for heterogeneity in treatment effects by setting or, alternatively, as testing for differences between treatment arms in a given setting. Suppose the sampling variance is equal across j and k for simplicity, i.e., $\sigma_j = \sigma_k = \sigma$. In this case, as detailed by Gelman and Tuerlinckx (2000), a classical test would call an observed difference significant at $p < 0.05$ if:

$$|Y_j - Y_k| > 1.96\sqrt{2}\sigma \tag{19}$$

If a Bayesian were to construct a 95% confidence interval for $\theta_j - \theta_k$, however, this interval would be represented by $\widehat{\theta}_j - \widehat{\theta}_k \pm 1.96\sqrt{V_j + V_k}$, where V_j is the variance of $\theta_j|\mu, \tau, Y$. It will later be observed that $V_j = 1/(1/\sigma_j^2 + 1/\tau^2)$, and the Bayesian analog to the classical test would be:

$$|Y_j - Y_k| > 1.96\sqrt{2}\sigma\sqrt{\frac{\tau^2 + \sigma^2}{\tau^2}} \tag{20}$$

This example illustrates that a measure analogous to I^2 is important in discerning heterogeneity across studies.[8]

8. As the number of studies increases, s^2 approaches σ^2 assuming a common σ.

Gelman and Tuerlinckx (2000) use this framework to examine what they call Type S and Type M errors, further discussed in Gelman and Carlin (2014). For Gelman and Tuerlinckx (2000), a Type S (sign) error is the probability that a claim is made that $\theta_j > \theta_k$ when in reality $\theta_j < \theta_k$.[9] A Type M (magnitude) error can be interpreted as an exaggeration factor, i.e., the expected value of a replication effect divided by the hypothesized true effect. Gelman and Carlin (2014) consider both types of error only for those results that were statistically significant and focus on predicting the effect of a replication in the same setting.

I argue that a similar approach can be taken to consider generalizability to another setting. Instead of Gelman and Tuerlinckx's Type S error, a policymaker may care about the probability that $\theta_j < 0$ when $\widehat{\theta}_j > 0$ or $\theta_j > 0$ when $\widehat{\theta}_j < 0$, regardless of the statistical significance of the estimate of θ_j. Similarly, analogous to the Type M error, a policymaker may care about the MSE of an estimate of θ_j, and this MSE can be predicted given estimates of τ^2.

In summary, τ^2 and I^2 are intrinsically related to the problem of how one might interpret evidence from a particular study.

2.6. Using Heterogeneity Measures to Extrapolate

With this background in place, I will now tie together the heterogeneity measures used in the literature and generalizability. Given a population P of potential studies, I will define generalizability as the ability to draw correct inferences from a set of studies, S, about a study j.[10] The inferences I will focus on are about the sign and magnitude of θ_j, answering the two questions posed in the introduction, namely 1) given a particular set of studies, how likely are we to correctly guess the sign of the true effect of a similar study in another context?, and 2) by what magnitude is our prediction likely to be wrong?

These are not the only potential questions of interest when thinking about extrapolating from a set of studies. For example, one might want to know the

9. Gelman and Tuerlinckx (2000) consider a claim made $\theta_j > \theta_k$ if the estimate of θ_j is significantly greater than the estimate of θ_k.

10. Importantly, the target study, j, may or may not be in the same setting as any study in S and it need not even share the same implementation details. However, θ_i and θ_j should be draws from the same distribution for any i in S; this enables any parameters estimated using S to be informative about θ_j. This distributional assumption matters when considering literature that are biased, such that the studies that were carried out were special in some way that affects their treatment estimates. Note, however, that I am explicitly not imposing that the true effects must be similar. If there are a variety of contexts and study-generating processes causing a wide dispersion of treatment effects, that should be captured in the τ^2 that is estimated using the studies in S. Instead, I only require the true effects to be from the same distribution.

More research into the biases introduced in the study-generating process is welcome. However, I regard the question of whether the results of many studies could be described as coming from the same distribution as an empirical question - part of the broader question of how well the model fits the data. Another possible way in which the model could be misspecified is if the error is not truly normally distributed. Any model misspecification can be considered empirically, in that one could fit a portion of the data to the model and use it to try to make predictions out of sample, though one would not be able to attribute the source of the error to e.g., research biases or other model misspecifications. I will consider this issue later in the paper.

probability that the true effect of a program in a given context falls above a certain non-zero threshold or that it falls within a given range. One may instead want to know the likelihood that a potential study will find a significant result assuming a given sample size and standard deviation. However, the two focal questions about sign and magnitude are certainly part of what one might care about.[11] I concentrate on these questions for clarity but note that the model can be used to answer many other questions.

From here, the approach is straightforward. First, consider the probability that one can accurately predict the sign of the true effect of a program in a given study j that is yet to be conducted.[12] The best guess as to the effect of a program in a new setting, without specifying a more complex model, is simply $\widehat{\mu}$, the best estimate of μ. Given values of μ, τ^2 and σ_j^2, one can calculate how likely it is that a correct inference will be made about the sign of the true effect of the program. Figure 1a plots curves representing particular constant probabilities assuming $\mu = 0.12$ (the mean standardized effect in the data) over various τ^2 and sampling variances. For simplicity in exposition I assume a common sampling variance across studies. Figure 1b similarly plots the average magnitude of the prediction error, again using standardized values. In practice, μ and τ^2 can be estimated using all N observations within each intervention-outcome combination, and for sufficiently high values of N one may think these estimated $\widehat{\mu}_N$ and $\widehat{\tau}_N^2$ are stable and approximate μ and τ^2. The sampling variance was assumed known in generating these simulated curves.[13] The figures are overlaid with triangles with indicative values for the 57 intervention-outcome combinations used in this paper. For these markers, estimates of τ are based on all N observations in each intervention-outcome combination, and Higgins and Thompson's approximation of a common sampling error, s, is assumed to approximate σ (2002).[14] These figures show that both the sampling variance and the true inter-study variation, τ^2, are important for making correct inferences about another study, but the two also interact: for large σ, decreasing τ^2 will not lead to large improvements in the accuracy of one's inferences. The intuition is that if the data are sufficiently noisy, whether the true effects vary from

11. For instance, a policymaker may prefer positive results to negative results for political reasons; they may also imagine beneficiaries might more strongly dislike a given harm than appreciate a comparable benefit. Then whether a potential program was likely to have a positive effect in their context would be important. The magnitude of the likely effect of the program is also something that would naturally enter into one's evaluation of the benefits of a particular program, and researchers conducting power calculations would also like to know the likely effect of a program in a given setting. Thus, error in predicting the magnitude of these effects is important.

12. This exposition will refer to results of a study. One may think that a study's results will be a function of the exact program variant and the context. At this stage, I do not have to model a study's results more explicitly. In particular, a program may vary in implementation or content from one study to another, so long as this variance can be estimated.

13. In reality, each study has its own sampling variance and the literature uses the standard error of a point estimate to estimate it.

14. This is because each study within an intervention-outcome combination has a different sampling variance, so some aggregate measure must be created. I only use s when a common measure is required, i.e., only in this figure and in estimates of I^2, never in the estimation of any θ_i, μ, or τ^2.

study to study will not be pivotal in making the correct prediction. On the other hand, reducing the sampling variance will always help, even in cases of large τ^2.

The above exposition focused on a random-effects model for the sake of clarity, without considering any potential explanatory variables. If one could leverage other data to build a better model, one may be able to obtain better estimates. Later in this paper, I will leverage mixed models that do seek to explain some of the observed variation.

Figure 2 shows how the two aforementioned questions about sign and magnitude would be answered by a random-effects model if applied to one illustrative intervention-outcome combination in the data (the effect of conditional cash transfers on school enrollment rates). In particular, to create each point, results from n studies relating to the effect of conditional cash transfers on enrollment rates (point estimate and standard error) are independently drawn and a best estimate of θ_{n+1}, $\widehat{\theta}_{n+1}$, is formed. Since this is a random-effects model, $\widehat{\theta}_{n+1}$ is simply $\widehat{\mu}_{n+1}$. Then this estimate of θ_{n+1} is compared to a draw of θ_i generated from $\theta_i \sim N(\widehat{\mu}_N, \widehat{\tau}^2_N)$, where $\widehat{\mu}_N$ and $\widehat{\tau}^2_N$ are the estimates of μ and τ^2 that are obtained from the random-effects data using all data for that intervention-outcome combination, assuming they approximate the true underlying parameter values μ and τ^2.[15] Similar figures for all intervention-outcomes in the data are provided in an appendix. In these figures, the predictions do not improve by much after the first few studies, an important point that will be discussed more later.

These figures assume the Bayesian model is true. I can empirically examine how well the model fits the data. While Figure 2 compares the estimates of $\widehat{\theta}_{n+1}$ to values of θ_i drawn from $\theta_i \sim N(\widehat{\mu}_N, \widehat{\tau}^2_N)$, one may instead wish to be more agnostic as to whether the model is correctly specified and compare the predictions of θ_{n+1} to a draw of Y_i.[16] These figures are provided in Online Appendix E. The figures making a comparison to $\widehat{\theta}_i$ and Y_i often have similar probabilities of making the correct inference about the sign of a similar study or its magnitude. Those using $\widehat{\theta}_i$ do not necessarily do better or worse than those using Y_i. When real data are used, however, drawn without replacement, it is no longer the case that the probability of making the correct inference about the sign of a similar study or its magnitude monotonically improves with the number of data points used.[17] Later in the paper, I will perform

15. The figures do not show a monotonic relationship between the accuracy of the estimate and the number of studies used because the studies can have quite different values from one another so the prediction error can be quite large at times. The studies are drawn in a random order in each simulation and increasing the number of simulations helps to average this out across simulations.

16. Yet another alternative would compare $\widehat{\theta}_{n+1}$ to a value of θ_i estimated from a draw of $\boldsymbol{Y_i}$ as in Equation 18. One might think that this estimated θ_i would capture the true effect of the study better than $\boldsymbol{Y_i}$, due to pooling. Nonetheless, I focus here on comparisons with $\boldsymbol{Y_i}$ to remain agnostic as to the appropriateness of the model.

17. An example can clarify why this is the case. Consider an intervention-outcome combination which has three point estimates: two small, insignificant and negative ones and one large, precisely-estimated positive one. If the two negative data points are drawn first, and used to generate the estimate of the third, they will mispredict the sign of the last, positive data point. If a negative data point and the positive data point are drawn first, they will also mispredict the remaining negative data point. Hence, the probability of

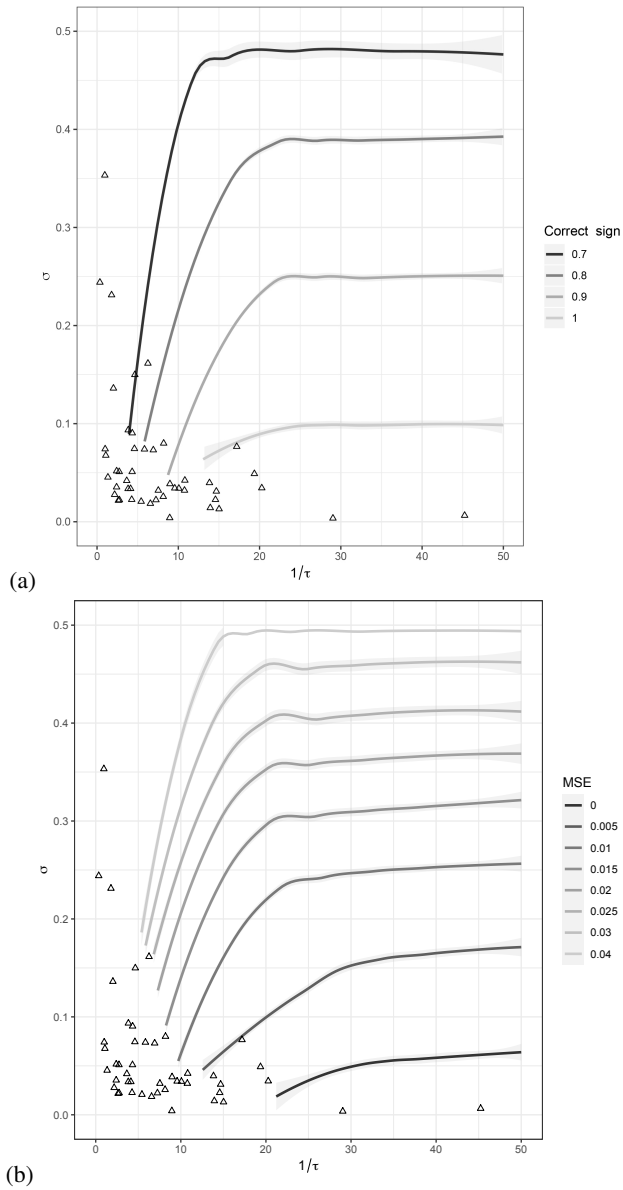

FIGURE 1. Heterogeneity measures and extrapolation. The top figure plots the probability of making the correct inference about the sign of the underlying parameter θ_j of some new study, j, assuming a mean standardized effect size of $\mu = 0.12$ and that τ^2 and σ^2 are known. The bottom figure plots the MSE of the prediction of the magnitude of θ_j, under the same assumptions. The range of values for τ and σ plotted here was chosen because these represent common estimated values in the data; the overlaid triangles represent intervention-outcome combinations in the data, using standardized values. For each intervention-outcome combination, τ is estimated using a random-effects model; to estimate a σ held in common across the intervention-outcome combinations, I use Higgins and Thompson's approximation, s.

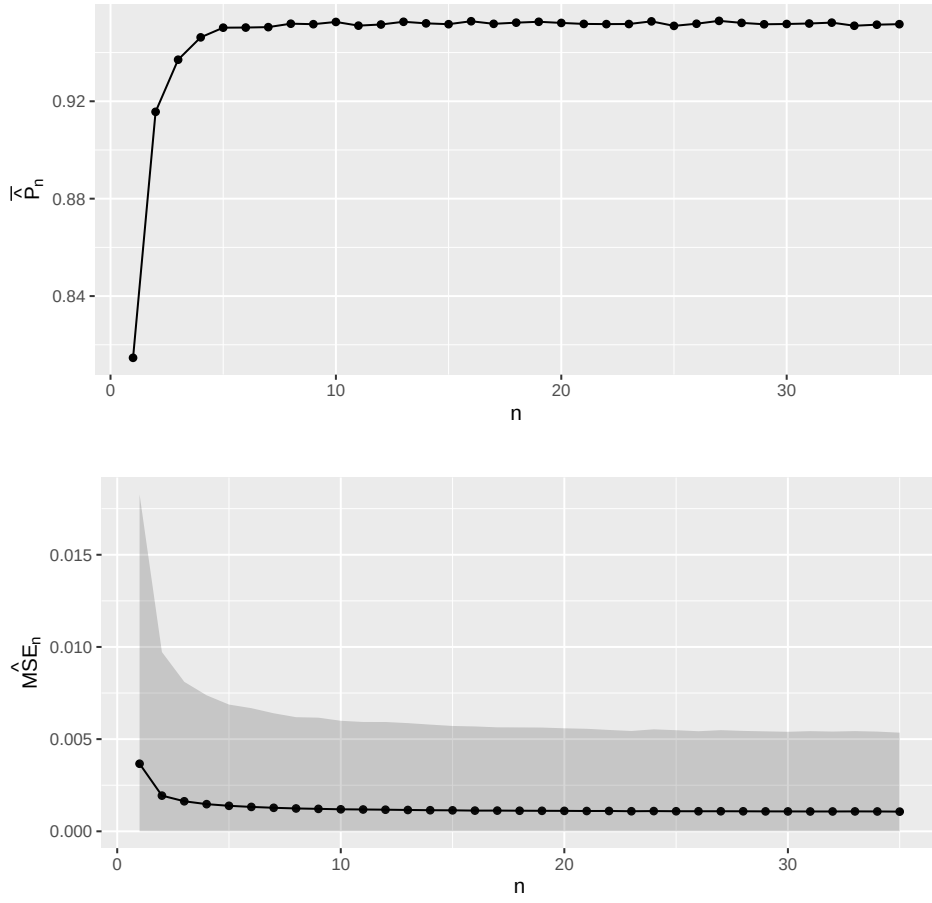

FIGURE 2. Example predicting the effects of a conditional cash transfer program on enrollment rates. These figures plot: (1) the probability of making the correct inference about the sign of the underlying parameter θ_{n+1} of some new study, given a certain number of studies, n, with estimates with which to make that guess; (2) the MSE of the best guess of θ_{n+1}.
n represents the number of studies used to form the estimate; for each intervention-outcome combination the maximum n used is $N-1$, so as to leave something to predict. 100,000 simulations are used and the mean probability of making the correct inference about the sign and the MSE is calculated for each n as described in the text. In the bottom part of the figure, the black line indicates the mean MSE; the 95% interval is provided by the shaded area. In the top part of the figure, the black line represents the mean probability of making the correct inference about the sign, but a 95% interval is not added since for any one given run the outcome will be 0 (incorrectly predicted the sign) or 1 (correctly predicted the sign).

additional checks to gauge the fit of the model.

correctly predicting the sign of the last study is actually zero when the maximum number of data points are used, no matter how many simulations are run.

Again, it should be noted that these are very simple models with no explanatory variables. A more complicated mixed model should obtain even better results, although in this particular example the probability of making the correct inference about the sign of θ_j is already quite high and the MSE quite low.[18]

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. Its main focus is on gathering results of impact evaluations and analyzing them through meta-analysis. AidGrade began 10 meta-analyses each in 2012 and 2013, and it followed the standard stages that are part of any Cochrane review:[19] topic selection; a search for relevant papers; screening of papers; data extraction; and data analysis.

In the sections that follow, I will briefly describe the main features of the data collection process, focusing on how interventions and studies were selected for inclusion, how variables were selected, and how results were extracted. A more detailed account is provided in an appendix. There were minor differences in the procedures followed for the meta-analyses began in 2012 and 2013; I will describe the process followed for those meta-analyses started in 2013 and note any differences in the process that pertain to the meta-analyses begun in 2012.

3.1. Selection of Interventions

Four AidGrade staff members each independently made a preliminary list of interventions, which were then combined. In 2012, there was no staff and this list was made by the author. Pilot searches were conducted using SciVerse and Google Scholar to determine whether there might be enough studies for a meta-analysis.[20] These pilot searches shortlisted 12 interventions in 2012, and 42 interventions in 2013. These were posted on AidGrade's website in 2013 and the general public was asked to vote on interventions they wanted covered. Respondents could select up to three interventions from those on the shortlist, and a space was provided for adding an "other" option. In the eight-day voting window, 452 votes were cast by 158 individuals, with 20 votes cast for the "other" option. The same procedure was followed in 2012, but the public vote did not influence the interventions ultimately selected, as it was discovered that two of the 12 interventions posted on the website that year did not have any outcome variables in common, which would preclude any meta-analysis. Thus, the other 10 interventions

18. This is a function of conditional cash transfer programs typically having small positive effects on enrollment rates.

19. The interested reader is referred to Part 2 of the current Cochrane Handbook for Systematic Reviews of Interventions by Higgins and Green (version 5.1 2011).

20. At this stage, the pilot searches only needed to identify two papers for an intervention to not be rejected.

with outcome variables in common were selected.

The outcome variables were categorized into three types, defined as follows. First, a set of "strict" outcomes captured results that measured the exact same thing, e.g., height in centimeters. "Loose" outcomes were those that measured the same variable but were defined in slightly different ways across studies. For example, different papers might use different hemoglobin threshold values for anemia. Finally, a set of "broad" outcomes was added retroactively to capture whether the outcome was an "economic", "educational", or "health" outcome. As it was important for the meta-analyses to compare similar outcomes, a rule was set that after searching for and screening papers for inclusion, the identified papers would be checked for "strict" outcomes held in common; if at least three papers on a common outcome were not found, that intervention would be replaced.

In 2013, from the shortlist of 42 interventions only 10 could be covered due to capacity constraints. These were selected partially through randomization to ensure balance between the included and excluded topics. However, the winner of the public vote, women's empowerment programs, was automatically selected to be included in the set of meta-analyses. In this process, interventions were matched prior to randomization using nearest neighbor matching.[21] After randomly restricting attention to half the topics, those covered by the most papers in the pilot searches were selected for inclusion, based on having found that many interventions from the meta-analyses begun in 2012 had relatively few papers sharing outcome variables. However, some interventions were still only covered by three studies. The interventions selected in each round of meta-analysis are listed in Table 2.

3.2. Identification of Papers

A comprehensive literature search was done using the search aggregators SciVerse, Google Scholar, and EBSCO/PubMed. The online databases of the Abdul Latif Jameel Poverty Action Lab, Innovations for Poverty Action, the Center for Effective Global Action and the International Initiative for Impact Evaluation were also searched. Finally, the references of all existing meta-analyses or systematic reviews turned up by the search were reviewed for completeness.

21. To obtain balance among the interventions included and excluded, each shortlisted topic was matched with another of the shortlisted topics based on how many likely impact evaluations the pilot searches identified for each; how many votes they received in the public vote; the overall theme of the interventions (e.g., education, health) according to the database of an external organization, AidData, after matching the interventions to AidData activity codes; and the recent aid commitments for the intervention as reported in AidData's database. The theme had to match exactly within each pair. For each of the three other factors, each topic was assigned a score on an index between zero and one representing where it stood among the other interventions; the index took the value: (topic value - minimum value among topics) / (maximum value among topics - minimum value among topics). 32 topics were successfully matched in this way using nearest neighbor matching without replacement. The remaining unmatched topics were singletons under their respective themes. For example, if there were an odd number of health-related interventions, the last health-related intervention would be by itself after others were matched. These last topics were independently randomized.

TABLE 2. List of development programs covered.

2012	2013
Conditional cash transfers	Contract teachers
Deworming	Financial literacy
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

Notes: This table lists the development programs considered in this paper. 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 data used in this paper.

Any impact evaluation on a given intervention was included, except those in high-income countries.[22] Both published studies and working papers were included. The particulars of the search and inclusion criteria used for each intervention is available in an online appendix. Screening proceeded in steps with the title, then the abstract, and finally the full text screened.

3.3. Selection of Variables and Data Extraction

All data were entered independently by two different coders and any discrepancies were reconciled by a third. In total, apart from a field specifying the topic, 85 fields were coded for each paper, including 13 fields with identifying information (author, publication year, program name, etc.); these were converted to 89 variables; the full list of variables and the coding manual is available as an online appendix). Additional topic-specific variables were coded for some interventions, such as the median and mean loan size for microfinance programs. This paper focuses on the variables held in common across the interventions, except when a mixed model is used for several intervention-outcome combinations covered by a large number of studies. The common variables include general identifying information (such as author and publication year); methodological information (such as the identification strategy used, whether the study was randomized by cluster, and whether it was blinded), characteristics of the intervention (such as location, duration between the start of intervention to the start of midline or endline data collection, intervention implementer,

22. The World Bank (2015) country classification system was used for this, with "high-income" countries excluded.

and characteristics of the sample), whether the study reported key information in the paper text (such as attrition and study costs), and finally, the results themselves.

Several key decisions relating to the data collected are worth highlighting. First, there was little choice over the selection of the results variables, since these needed to capture the actual way that results were reported in a paper. For the variables capturing study characteristics there was more choice, and here it was thought important for the interpretation of the results to know more about the methods used and context of the study.

Since this paper pays particular attention to the program implementer, it is worth discussing in more detail how this variable was coded. Implementers could initially be coded as governments, NGOs, private sector firms, or academics. There was also a code for "other" or "unclear". It was ultimately decided to consider NGOs and academic research teams together as it turned out to be practically difficult to distinguish between them in the studies, especially as the papers frequently used passive voice (e.g., "X was done" without noting who did it).

Since this paper focuses on heterogeneity of impact evaluation results, I focus on the "strict" outcomes, defined previously as outcomes that measured the exact same thing. Analyzing studies with "strict" outcomes helps exclude those sources of variation that stem from different outcome measures being used.[23]

There were also several closely related "strict" outcome variables, such as diarrhea prevalence and diarrhea incidence, or enrollment rates and attendance rates.[24] I keep these outcomes separate because they do not follow the "strict" rule: they are not measuring the exact same thing, and one would consequently expect some natural variation in their results.

Studies tended to report results for multiple specifications. AidGrade focused on those results least likely to have been influenced by author choices, i.e., specifications with the fewest controls, apart from fixed effects. Where a study reported results using different methodologies, coders followed 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 collected both the aggregate results and any

23. The exceptions to this rule were that the impact of bed nets on malaria and the impact of micronutrients on anemia were considered despite malaria and anemia being "loose" outcomes, because these outcomes were typically among the primary goals of their respective interventions. Malaria was the unique outcome held in common across many studies of bed nets programs, and including anemia also results in fewer papers being discarded for not having outcome variables in common.

24. "Prevalence" measures capture the proportion of the population experiencing the disease or symptom at one point in time. "Incidence" measures instead capture the rate of occurrence of new cases of a disease or symptom over a period of time. It is important to distinguish between these, as they may differ substantially. For example, if an illness takes a long time to cure, shifts in its prevalence rate may not be easily apparent, whereas shifts in its incidence rate may be more rapidly observed. These outcomes also are reported in different ways.

results by subgroup.[25]

There may seem to be some tension between using the authors' preferred methodology where specified but also focusing on those results with the fewest controls. This approach was taken due to the belief that it would be much easier for researchers to consciously or subconsciously engage in specification searching by adding covariates or restricting attention to certain subgroups. In contrast, it may be harder to engage in specification searching by changing methodology. First, researchers tend to be rewarded for pursuing the most credible methods wherever possible, and so one might expect that where researchers have a choice they will always pick the most credible methods. Further, the method used was often implicitly selected before the beginning of the study; most of the studies in the database are randomized controlled trials, and these are usually planned in advance. There were few instances in which a paper reported results using two different methods.

3.4. Data Description

I focus on those papers that passed all screening stages in the meta-analyses. 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. In some cases, multiple papers by the same authors or multiple versions of the same paper reported results for the same outcomes; as these were correlated, I used only the most recent result for each outcome in analysis. After removing these duplicates, the number of observations drops from 15,024 results collected across 635 papers to 1,932 results from 307 papers when restricting attention to only those results that can be compared with results from another paper on the same intervention-outcome. The implication is that even when papers report on common outcomes, those common outcomes represent a small subset of the results a paper reports.

These 1,932 results include multiple results from the same intervention-outcome-paper on different subgroups or over different time periods. For most of the analyses in this paper, I collapse the data to one observation per intervention-outcome-paper to avoid dependence between observations (Higgins and Green 2011). Where results had been reported for multiple subgroups (e.g., women and men), I aggregate them as in the Cochrane Handbook's Table 7.7.a. Where results were reported for multiple

25. There was one exception to this rule, which was if an author appeared to only be 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.

time periods (e.g., six months after the intervention and twelve months after the intervention), I use the most comparable time periods across papers. Sometimes, a paper provided more than one set of subgroups, such as results for girls and boys and, separately, results for three different age groups. When aggregating across different subgroups, I use the minimal set of subgroups that could be aggregated (i.e., girls and boys in the example). This minimal set was comprised of 887 results. Aggregating them reduced the number of results to 698 (across 307 papers) if using the "loose" outcomes and retaining those intervention-outcome combinations covered by at least two papers. For the outcomes considered in this paper (the "strict" outcomes plus the loose outcomes for malaria and anemia prevalence), this reduced the number of results to 646 (across 276 papers) if retaining those intervention-outcome combinations covered by at least two papers and 576 results (across 251 papers) if retaining those intervention-outcome combinations covered by at least three 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.

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 $\text{SMD} = (\mu_1 - \mu_2)/\sigma_p$, 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.[26] The signs of the results were also adjusted so that a positive effect size always represents an improvement. Data could not always be standardized, as the standard deviation of the outcome variable was often not reported. Thus, the standardized data consist of only 612 results if retaining those intervention-outcome combinations covered by at least two papers and 561 if retaining those intervention-outcome combinations covered by at least three papers.

Figure 3 shows the raw distribution of effects for each of the intervention-outcome combinations. This figure suggests a fair amount of variation. In general, interventions are rarely distinguishable from one another in terms of their effects on a particular outcome, with their effect sizes tending to overlap substantially. Table B.1 in Online Appendix B 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 only provides ambiguous evidence of

26. Ideally, the study would report σ_p, in which case that value was used. When it reported standard deviations separately for the control and treatment group, these were pooled using the formula in the Cochrane Handbook's Table 7.7.a. When these data were not available, the standard deviation in the control group was preferentially used, followed by the standard deviation in the treatment group, followed by the standard deviation of the outcome variable from other studies within the same intervention-outcome combination.

the typical effects in that intervention-outcome. Simply tallying the numbers in each category is known as "vote counting" and can be misleading.[27].

Table 3 further summarizes the distribution of papers across interventions and highlights the fact that papers do not frequently study the same outcomes. This is consistent with the story that researchers each want to publish one of the first papers on a topic. Figure B.2 in Online Appendix B disaggregates these numbers by intervention-outcome combination.

TABLE 3. Descriptive statistics: Distribution of strict 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	11	13	17
Financial literacy	3	4	5
HIV/AIDS education	5	3	4
Improved stoves	4	2	2
Insecticide-treated bed nets	1	10	10
Irrigation	2	2	2
Micro health insurance	3	2	2
Microfinance	6	4	5
Micronutrient supplementation	20	24	37
Mobile phone-based reminders	2	3	3
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	7	9
Women's empowerment programs	2	2	2
Average	4.6	6	8.2

Notes: This table shows the distribution of strict outcomes across interventions. As described in the text, two "loose" outcomes, malaria and anemia prevalence, are included due to their having frequently been among the primary goals of the intervention.

4. Results

The previous sections motivated the use of some measures of heterogeneity, explicitly linked them to generalizability through a Bayesian model, and described the data.

27. For example, if a review of the literature uncovered many papers with small sample sizes and insignificant effects, one might be tempted to conclude the intervention "didn't work" when it could merely be that each study was underpowered and if the results were pooled in a meta-analysis they would be significant. Many authors have described further unpalatable properties of vote counting (e.g., Koricheva and Gurevitch 2013; Combs et al. 2009; Hedges and Olkin 1980)

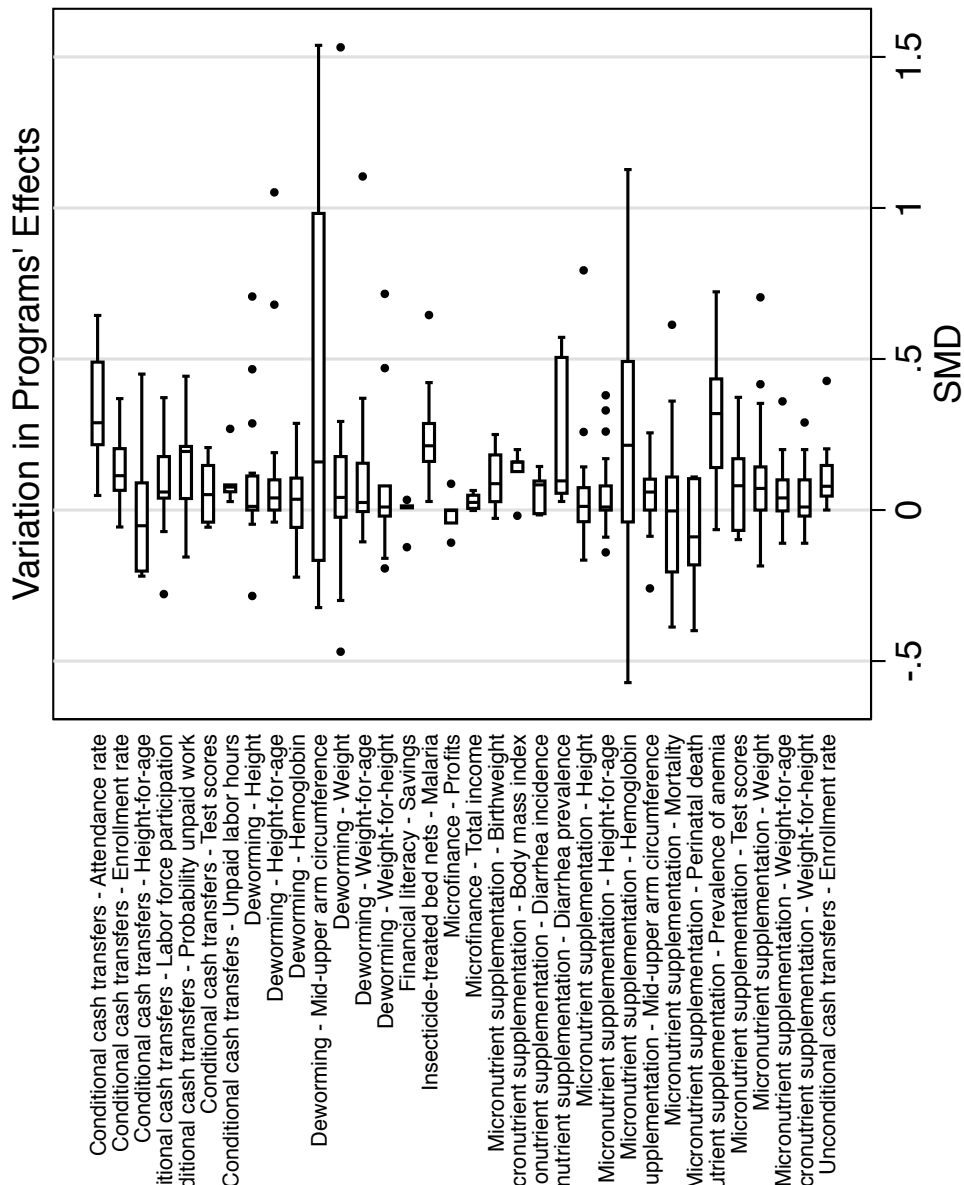

FIGURE 3. Dispersion of estimates. This figure provides a box-and-whisker plot of the effect sizes found for each intervention-outcome combination covered by at least 5 papers. The dots represent outliers (defined by convention as 1.5 IQR beyond the nearest quartile), the whiskers the remaining maximum and minimum values, and the boxes the interquartile range and median value. One observation with an effect size greater than 2 is omitted for legibility.

This section provides values for these measures and explores how they vary with study or intervention characteristics. The data set contains 96 intervention-outcome combinations, of which 57 are covered by at least three papers. The rest of this paper will focus on this set of 57 intervention-outcomes, unless otherwise specified.

4.1. Without Modeling Heterogeneity

Table 4 presents estimates of the likelihood of making a correct inference about the sign of a similar study, the expected $\sqrt{MSE}$, τ^2 and I^2 for each intervention-outcome combination. For reference, estimates of μ are also provided, along with $\tau/|\mu|$ and the "typical" σ_i among studies in an intervention-outcome, using Higgins and Thompson's aggregation, s. Unstandardized values are used.

The median probability that the sign of a similar study would be correctly predicted for these intervention-outcome combinations was 61%. Those intervention-outcome combinations with the highest likelihood that a prediction about a similar study would have the right sign had the lowest values of $\widehat{\tau}_N/|\widehat{\mu}_N|$. Recall that the probability of making the correct inference about sign does not depend specifically on $\tau/|\mu|$, but it does depend on both τ^2 and μ, and the ratio can help in interpreting τ^2. The $\sqrt{MSE}$ may likewise be easiest to interpret relative to $\widehat{\mu}_N$. The median $\sqrt{MSE}/|\widehat{\mu}|$ for these intervention-outcome combinations was 2.49. In other words, a prediction of a result in a new setting is likely to be wrong by about 249% unless some of the variation can be explained by a model.

Some of the lowest values of $\widehat{\tau}_N/|\widehat{\mu}_N|$ are for conditional cash transfers and health-related interventions such as the impact of bed nets on malaria. Among those with the highest $\widehat{\tau}_N/|\widehat{\mu}_N|$ are the financial interventions, i.e., microfinance and financial literacy training. For only a few intervention-outcome combinations can one make the correct inference about the sign of a similar study at least 90% of the time: bed nets reliably decrease malaria and CCTs improve enrollment rates. For microfinance and financial literacy, the probability of making the correct inference about the sign of a similar study was only slightly better than 50% in most cases. That said, within a given intervention, the probability of making the correct inference about the sign of a similar study varied by outcome. For example, for conditional cash transfers, the probability of making the correct inference about height-for-age was only 51%, likely a function of the estimated average effect size being so small. It should also be noted that those interventions covering a large number of outcomes would be more likely to be represented at both the high and low end of the $\widehat{\tau}_N/|\widehat{\mu}_N|$ spectrum just by chance.

Table 5 further summarizes these results by creating three bins for $|\widehat{\mu}_N|$ and $\widehat{\tau}^2_N$ such that a third of the intervention-outcomes fall into each bin and then reporting the mean correct rate and $\sqrt{MSE}$ for those intervention-outcomes that fall in each of the cells of the resultant 3x3 table. This table shows that for intervention-outcome combinations with a low $|\widehat{\mu}_N|$ and medium or high $\widehat{\tau}^2_N$, as well as those with a medium $|\widehat{\mu}_N|$ and high $\widehat{\tau}^2_N$, the sign of another study can be predicted with just better than a 50% chance. For reference, the cutoff thresholds for "low" and "high" $|\widehat{\mu}_N|$ were

0.057 and 0.167, and for $\widehat{\tau}^2_N$ these thresholds were 0.030 and 0.175, respectively.[28] Intervention-outcomes with larger $\widehat{\tau}^2_N$ are likely to have larger $|\widehat{\mu}_N|$, but not all do. The right hand side of the table provides the number of intervention-outcomes that fall into each cell, suggesting that some of the summary statistics be treated with caution due to the small number of intervention-outcomes involved. Still, the table may be helpful in summarizing the disaggregated results.[29]

It should be noted that τ^2 depends on how narrowly an intervention-outcome is defined. If outcomes were defined more broadly, for example, τ^2 would appear to rise. By using narrowly defined outcomes, i.e., "strict" outcomes plus malaria and anemia prevalence, these results err on the side of smaller estimated τ^2.[30] Using "strict" outcomes reduces the number of studies that can be included in the analysis, but it minimizes τ^2 and yields better predictions than increasing the number of studies in a cell for these data. As previously observed, the marginal benefits to prediction of including an additional study fall precipitously over the first few studies. Also, it is helpful to be able to distinguish between different potential sources of variation for interpretability.

4.1.1. 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". Here, I consider two measures.[31]

First, I 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

28. It is hard to convert these to values of $\widehat{\tau}_N/|\widehat{\mu}_N|$, given that the lower $\widehat{\tau}_N$ within a cell, the lower $\widehat{\tau}_N/|\widehat{\mu}_N|$, and the lower $|\widehat{\mu}_N|$ within a cell, the higher $\widehat{\tau}_N/|\widehat{\mu}_N|$, but if one were at exactly the cutoff threshold for "low" $\widehat{\tau}_N$ and "low" $|\widehat{\mu}_N|$, this would correspond to a $\widehat{\tau}_N/|\widehat{\mu}_N|$ value of 3.04; at the "high" cutoff threshold for $|\widehat{\mu}_N|$, the "low" $\widehat{\tau}_N$ cutoff corresponds to a equate to a $\widehat{\tau}_N/|\widehat{\mu}_N|$ value of 1.04. At the "high" $\widehat{\tau}_N$ cutoff, the "low" and "high" cutoffs for $|\widehat{\mu}_N|$ yield $\widehat{\tau}_N/|\widehat{\mu}_N|$ values of 7.34 and 2.50, respectively. Again, there will be great variation in $\widehat{\tau}_N/|\widehat{\mu}_N|$ within a cell depending on the exact values taken by $\widehat{\tau}_N$ and $|\widehat{\mu}_N|$.

29. Again, it should be noted that studies on some interventions reported more outcomes than others. Due to this fact and the possibility of unmodeled correlation between different outcomes, this table should not be interpreted as providing low, medium and high values of $|\widehat{\mu}_N|$ and $\widehat{\tau}^2_N$ for interventions.

30. As discussed, the intervention-outcome combination of bed net programs - malaria had the lowest $\widehat{\tau}_N/|\widehat{\mu}_N|$ and was associated with the highest probability of making the correct inference about the sign of a similar study. Anemia prevalence also fared well along these measures. Hence, the choice to include these outcomes does not appear to have biased the overall results to make studies appear more heterogeneous.

31. There are also other ways to measure paper quality. I 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.

interventions being tested are not typically suited to blinding, I consider all papers scoring at least a 3 to be of "high quality".

In an alternative specification, I 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 companion paper finds that RCTs exhibit the fewest signs of specification searching and publication bias (Vivalt 2019). It should be emphasized that without building an explicit model for potential biases, I would have no way of separating these biases from true, underlying heterogeneity in treatment effects. Thus, looking at only studies that were RCTs and hence less subject to specification searching and publication bias provides a good robustness check.

TABLE 4. Heterogeneity measures for treatment effects within intervention-outcomes.

Intervention	Outcome	Units	$\widehat{P(Sign)}$	$\widehat{\sqrt{MSE}}$	$\hat{\tau}^2_N$	$\hat{I}^2_N$	$\frac{\hat{\tau}_N}{\lvert\hat{\mu}_N\rvert}$	$\hat{\mu}_N$	$\hat{s}_N$	N
Conditional Cash Transfers	Retention rate	percentage points	0.65	0.01	0.000	0.86	1.51	-0.01	0.00	5
Conditional Cash Transfers	Attendance rate	percentage points	0.76	0.07	0.001	0.80	0.57	0.05	0.02	14
Conditional Cash Transfers	Labor force participation	percentage points	0.77	0.03	0.001	0.92	1.33	-0.02	0.01	18
Unconditional Cash Transfers	Enrollment rate	percentage points	0.87	0.03	0.001	0.90	0.86	0.04	0.01	13
Conditional Cash Transfers	Enrollment rate	percentage points	0.95	0.03	0.001	0.96	0.60	0.05	0.01	36
Financial Literacy	Has savings	percentage points	0.64	0.05	0.001	0.61	1.48	0.02	0.03	4
Micronutrients	Birthweight	kg	0.79	0.05	0.002	0.89	1.17	0.04	0.02	7
Rural Electrification	Enrollment rate	percentage points	0.79	0.09	0.002	0.65	0.69	0.07	0.04	3
Deworming	Hemoglobin	g/dL	0.54	0.08	0.004	0.56	3.71	0.02	0.06	14
Micronutrients	Weight-for-height	standard deviations	0.70	0.07	0.005	0.77	1.80	0.04	0.04	26
Micronutrients	Weight-for-age	standard deviations	0.72	0.09	0.009	0.89	1.76	0.05	0.03	31
Micronutrients	Mid-upper arm circumference	cm	0.73	0.10	0.009	0.82	1.55	0.06	0.04	17
Micronutrients	Height-for-age	standard deviations	0.67	0.11	0.011	0.90	2.21	0.05	0.03	33
Micronutrients	Diarrhea incidence	log risk ratio	0.80	0.14	0.015	0.82	1.05	-0.11	0.06	7
Financial Literacy	Has taken loan	percentage points	0.50	0.15	0.016	0.93	10.14	0.01	0.03	4
HIV/AIDS Education	Used contraceptives	percentage points	0.61	0.18	0.023	0.93	1.93	0.08	0.04	4
Conditional Cash Transfers	Probability unpaid work	percentage points	0.56	0.18	0.024	0.98	3.03	-0.05	0.02	5
Conditional Cash Transfers	Height-for-age	standard deviations	0.51	0.21	0.029	0.84	18.90	-0.01	0.07	7
Bed Nets	Malaria	log risk ratio	0.98	0.20	0.030	0.69	0.46	-0.38	0.12	10
SMS Reminders	Appointment attendance rate	log risk ratio	0.78	0.22	0.031	0.92	1.02	0.17	0.05	3
Micronutrients	Test scores	standard deviations	0.65	0.20	0.034	0.99	2.16	0.09	0.02	9
Conditional Cash Transfers	Pregnancy rate	percentage points	0.52	0.24	0.038	0.98	6.51	-0.03	0.03	3
Micronutrients	Weight	kg	0.76	0.21	0.041	0.96	1.39	0.15	0.04	31
Contract Teachers	Test scores	standard deviations	0.71	0.29	0.054	0.95	1.23	0.19	0.05	3
Conditional Cash Transfers	Gave birth at healthcare facility	percentage points	0.52	0.29	0.055	0.94	4.36	0.05	0.06	3
Performance Pay	Test scores	standard deviations	0.60	0.30	0.059	0.98	2.03	0.12	0.03	3
Conditional Cash Transfers	Skilled attendant at delivery	percentage points	0.57	0.31	0.062	0.90	2.47	0.10	0.08	3
Conditional Cash Transfers	Test scores	standard deviations	0.54	0.31	0.069	0.98	3.11	0.08	0.03	5
Deworming	Weight-for-height	standard deviations	0.54	0.29	0.075	0.98	4.59	0.06	0.04	11
Micronutrients	Body mass index	kg/m^2	0.75	0.31	0.077	0.99	1.31	0.21	0.03	5
Micronutrients	Mortality	log risk ratio	0.52	0.33	0.083	0.50	6.32	-0.05	0.29	11
Scholarships	Enrollment rate	percentage points	0.55	0.40	0.111	1.00	2.95	0.11	0.02	3

Deworming	Height-for-age	standard deviations	0.65	0.38	0.132	1.00	2.25	0.16	0.02	14
Deworming	Weight-for-age	standard deviations	0.61	0.40	0.145	1.00	2.74	0.14	0.02	12
Micronutrients	Perinatal death	log risk ratio	0.56	0.45	0.151	0.69	3.18	0.12	0.26	6
Micronutrients	Diarrhea prevalence	log risk ratio	0.65	0.45	0.156	0.90	1.77	-0.22	0.13	6
School Meals	Test scores	standard deviations	0.50	0.54	0.170	0.98	8.91	0.05	0.05	3
Micronutrients	Prevalence of anemia	log risk ratio	0.89	0.44	0.175	0.87	0.80	-0.52	0.16	13
Deworming	Mid-upper arm circumference	cm	0.53	0.46	0.176	0.99	4.93	0.09	0.04	7
Deworming	Weight	kg	0.59	0.44	0.182	0.99	3.33	0.13	0.05	17
School Meals	Enrollment rate	percentage points	0.50	0.66	0.216	0.90	11.57	0.04	0.16	3
Micronutrients	Stunted	log risk ratio	0.51	0.60	0.228	0.89	6.70	-0.07	0.17	3
Deworming	Height	cm	0.53	0.51	0.229	0.95	5.41	0.09	0.11	16
Micronutrients	Hemoglobin	g/dL	0.72	0.49	0.235	0.99	1.70	0.29	0.04	37
Micronutrients	Height	cm	0.64	0.50	0.244	0.96	2.81	0.18	0.10	29
Water Treatment	Diarrhea prevalence	log rate ratio	0.77	0.57	0.279	0.96	1.29	-0.41	0.10	9
Water Treatment	Diarrhea incidence	log rate ratio	0.75	1.02	0.791	0.96	1.28	-0.69	0.17	5
Conditional Cash Transfers	Unpaid labor hours	hours/week	0.81	1.17	0.993	0.83	0.98	-1.02	0.45	5
Micronutrients	Stillbirth	log risk ratio	0.51	1.19	1.023	0.85	8.10	0.12	0.42	4
Water Treatment	Dysentery incidence	log rate ratio	0.59	2.22	3.305	0.97	2.08	-0.88	0.31	3
Conditional Cash Transfers	Labor hours	hours/week	0.73	2.60	5.491	0.97	1.44	-1.63	0.42	7
Rural Electrification	Study time	hours/day	0.57	3.89	9.991	0.99	2.35	1.34	0.32	3
Financial Literacy	Savings	current US$	0.56	56.84	1100.337	0.92	1.79	18.58	9.71	5
Microfinance	Total income	current US$	0.59	65.55	2806.259	0.96	2.14	24.74	10.83	5
Microfinance	Profits	current US$	0.50	161.64	18134.689	0.96	22.66	5.94	28.31	5
Microfinance	Savings	current US$	0.50	211.67	29058.289	1.00	8.67	19.65	6.02	3
Microfinance	Assets	current US$	0.51	330.21	76265.430	0.99	5.40	51.17	28.59	4

Notes: $\widehat{P(Sign)}$ is the average estimated probability of making the correct inference about the sign of a particular true effect, θ_j, given all data in that intervention-outcome combination, and $\widehat{\sqrt{MSE}}$ represents the average estimated square root of the mean squared error of that prediction. $\widehat{\tau}^2_N$, $\widehat{I}^2_N$, $\widehat{\tau}_N/|\widehat{\mu}_N|$ and $\widehat{\mu}_N$ likewise present the average estimate for each parameter. $\widehat{s}_N$ estimates a common sampling error for each intervention-outcome using Higgins and Thompson's approximation. It is important in estimating $\widehat{I}^2_N$ and it provides a way to summarize the σ_i within an intervention-outcome combination, given they vary by study. However, the individual study-specific estimates of the sampling variance, σ_i^2, were used to generate the estimates of μ and τ and hence the other columns in the table. Each measure is calculated separately by intervention-outcome combination, without pooling across intervention-outcomes. Unstandardized values are used throughout. 10,000 simulations are run to calculate the probability of making the correct inference about the sign of θ_j and the MSE for each intervention-outcome combination. Wherever $\widehat{I}^2_N$ appears equal to 1.00, this is the result of rounding. This table reports results for all 57 intervention-outcome combinations covered by at least three studies.

TABLE 5. Summary of generalizability measures by heterogeneity measures.

	$\widehat{P(Sign)}$			$\widehat{\sqrt{MSE}}$			N		
	$\hat{\tau}^2_N$			$\hat{\tau}^2_N$			$\hat{\tau}^2_N$		
$\lvert\hat{\mu}_N\rvert$	Low	Medium	High	Low	Medium	High	Low	Medium	High
Low	0.688	0.515	0.500	0.08	0.35	0.66	14	4	1
Medium	0.733	0.603	0.534	0.13	0.33	0.64	4	10	5
High	0.980	0.756	0.634	0.20	0.34	64.49	1	5	13

Notes: This table summarizes the information provided in Table 4 by splitting the intervention-outcome combinations into three equal-sized groups according to $\lvert\hat{\mu}_N\rvert$ and $\hat{\tau}^2_N$ and then calculating the average value of $\widehat{P(Sign)}$ and $\widehat{\sqrt{MSE}}$ for the intervention-outcome combinations that fall in each cell. Note that since $\lvert\hat{\mu}_N\rvert$ tends to increase with $\hat{\tau}^2_N$, there are relatively few observations in some cells.

Table 6 provides summary measures of heterogeneity using the data that meet these two quality criteria.[32] The summary heterogeneity measures are not substantially different using these data. The common sampling error that is estimated, $\hat{s}_N$, appears slightly lower in magnitude for the 50th and 75th percentile for these studies. However, a t-test fails to reject that the mean of $\hat{s}_N$ among either group of results meeting these quality criteria is lower than the mean of $\hat{s}_N$ among all results, and this is true even when restricting attention to the half of each sample with the highest $\hat{s}_N$. These differences are minute enough to not translate into improvements in $\widehat{P(Sign)}$ or $\widehat{\sqrt{MSE}}$.

4.1.2. Model Checking. It is good practice to check the fit of the model using posterior predictive checks. These checks compare the data with the posterior distribution, under the intuition that for a model that fits the data well, the data should look similar to draws from the posterior distribution. To conduct a posterior predictive check, one takes some function of the data that is of interest and generates a test statistic, T, for that function using the data and the simulated posterior distribution. One then computes the probability that the test statistic in the posterior distribution is larger than that in the observed data. This defines the Bayesian p-value:

$$p = P(T(Y^{rep}, \theta) \geq T(Y, \theta)|Y) \tag{21}$$

where Y^{rep} can be thought of as the data that the model would predict a replication of all the studies would find, θ represents all the parameters in the model, including the hyperparameters, and Y represents the observed data.

These tests have a nuanced interpretation, and it is not the case that if a model fails to fit the data in some way it is necessarily a bad model. In particular, in the context of this paper, one might think that outliers in the original data that were based on small sample sizes are not great estimates of their respective true effect sizes θ_i

32. Full tables are provided in Tables B.2-B.3 in Online Appendix B.

TABLE 6. Heterogeneity measures by study quality.

	$\widehat{P(Sign)}$	$\widehat{\sqrt{MSE}}$	$\hat{\tau}^2_N$	$\hat{I}^2_N$	$\frac{\hat{\tau}_N}{\lvert\hat{\mu}_N\rvert}$	$\hat{\mu}_N$	$\hat{s}_N$	N
All studies								
25th percentile	0.54	0.15	0.016	0.87	1.33	-0.01	0.03	4
50th percentile	0.61	0.31	0.075	0.94	2.14	0.05	0.05	6
75th percentile	0.75	0.54	0.229	0.98	4.36	0.13	0.16	13
RCTs								
25th percentile	0.55	0.11	0.011	0.88	1.30	-0.04	0.03	4
50th percentile	0.65	0.33	0.075	0.95	1.97	0.05	0.05	7
75th percentile	0.74	0.50	0.224	0.98	3.58	0.13	0.12	14
Higher-quality studies								
25th percentile	0.55	0.14	0.015	0.89	1.47	-0.07	0.03	4
50th percentile	0.65	0.37	0.087	0.95	1.86	0.05	0.04	7
75th percentile	0.72	0.52	0.226	0.98	3.48	0.14	0.12	14

Notes: This table shows quantiles of the heterogeneity measures for different subgroups of studies: all studies, RCTs, and those studies considered "higher-quality" using the Jadad scale. As in Table 4, $\widehat{P(Sign)}$ is the average estimated probability of making the correct inference about the sign of a particular true effect, θ_j, given all data in that intervention-outcome combination, and $\widehat{\sqrt{MSE}}$ represents the average estimated square root of the mean squared error of that prediction. $\hat{\tau}^2_N$, $\hat{I}^2_N$, $\hat{\tau}_N/|\hat{\mu}_N|$ and $\hat{\mu}_N$ likewise present the average estimate for each parameter, and $\hat{s}_N$ estimates a common sampling error for each intervention-outcome using Higgins and Thompson's approximation. Each measure is calculated separately by intervention-outcome combination, without pooling across intervention-outcomes. Unstandardized values are used throughout. 10,000 simulations are run to calculate the probability of making the correct inference about the sign of θ_j and the MSE for each intervention-outcome combination. Wherever $\hat{I}^2_N$ appears equal to 1.00, this is the result of rounding. This table reports results for those intervention-outcome combinations covered by at least three studies.

but should rightly be thought to be closer to the mean within an intervention-outcome combination. In that case, some differences between the original data and the posterior distribution would be expected and even desired.

There are many test statistics that could be used to check whether the top-level assumption of normality is reasonable in my data. Following Bandiera et al. (2016), I check whether the center part of my data and the posteriors appear equally symmetrically distributed. Often a fairly large interval of the data is used for this kind of analysis, but some of my intervention-outcome combinations are small and it makes little sense to talk about a 10th percentile, for example, of an intervention-outcome combination with three studies. I thus start by considering the 25-75th percentile for all intervention-outcome combinations. For those intervention-outcomes with at least nine studies, I also consider the 10-90th percentile, and for those with at least 19 studies I additionally consider the 5-95th percentile. Taking the example of a test of symmetry over the 10-90th percentile in an intervention-outcome with nine studies, the test statistic would be written as follows:

$$T(Y,\theta) = |Y_{(9)} - \mu| - |Y_{(1)} - \mu| \tag{22}$$

where the ninth and first order statistics represent the 90th and 10th percentiles of the distribution and μ is the best guess of any data point in Y in a random-effects model. $T(Y, \theta)$ would be distributed around 0 if symmetric. The exact order statistic used depends on the intervention-outcome, given the different number of studies they contain and the different percentiles tested. $T(Y^{rep}, \theta)$ would similarly be written as:

$$T(Y^{rep}, \theta) = |Y^{rep}_{(9)} - \mu| - |Y^{rep}_{(1)} - \mu| \tag{23}$$

To calculate the p-value in Equation 20, I draw 1,000 values of μ and τ^2 from their posterior distributions and then draw Y^{rep}, using σ_i^2 drawn from the set of observed σ_i^2.

Table B.4 presents results. For none of the 57 intervention-outcome combinations are the observed data and the simulated replication data significantly different from each other in the 25-75th interval.[33] For one of the 22 intervention-outcome combinations with at least nine studies, the observed data and simulated replication data are significantly different in the 10-90th interval. For two of the seven intervention-outcome combinations with at least 19 studies, the observed data and simulated replication data are significantly different in the 5-95th interval. The larger the number of studies, the more likely the tails are to be skewed. The one intervention-outcome which failed the 10-90th interval test was the effect of CCTs on enrollment rates, an intervention-outcome with among the largest number of studies. This intervention-outcome combination also failed the test using the 5-95th interval, along with the effect of micronutrients on height. It should not be surprising that intervention-outcome combinations with a larger number of studies should fail these tests more often as it is easier to detect a misspecified model with more data points. Overall, the tests are encouraging, though the model may fit less well in the extreme tails.

4.2. Modeling Heterogeneity

If the observed heterogeneity in outcomes can be systematically modeled, one could make better predictions. I first look across different intervention-outcome combinations to examine whether effect sizes, $\hat{\tau}^2$ or $\hat{I}^2$ are correlated with any study or intervention characteristics. I then turn to look within a few specific intervention-outcome combinations and build a mixed model to try to explain the variance.

4.2.1. Across Intervention-Outcomes. Table 7 presents the results of a simple OLS regression of effect sizes on study characteristics, using standardized values.[34] Data for

33. The relevant thresholds are $p < 0.025$ and $p > 0.975$, given that for these tests a very high p-value also indicates a poor fit; to gauge fit one should test not just whether $T(Y^{rep}, \theta) \geq T(Y, \theta)$ but also whether $T(Y^{rep}, \theta) \leq T(Y, \theta)$.

34. Variables one might wish to include in this kind of regression and for which the data are not too sparse include: number of authors; publication year; publication code i.e., published or unpublished and type of journal; organization code; method; whether the study was blinded; country (aggregated here to region following the World Bank's geographic divisions to avoid including too many dummy variables);

10 of the 57 intervention-outcomes could not be standardized and hence are excluded from this table.[35]

I find that studies based on a smaller number of observations have greater effect sizes. This is what one would expect if specification searching were easier in small data sets. This pattern of results would also arise 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 were less well-targeted. Interestingly, government-implemented programs have lower effect sizes even controlling for sample size, compared to programs implemented by the private sector. Studies in the Middle East / North Africa (MENA) region may appear to perform slightly better than those in Sub-Saharan Africa (the excluded region category), but very few studies were conducted in MENA countries, so not much weight should be put on this. RCTs do not exhibit significantly different results than quasi-experimental studies within an intervention-outcome combination.

These regressions include intervention-outcome fixed effects so as to better isolate the variation in effect sizes that can be explained by study characteristics even across intervention-outcome combinations, and standard errors are also clustered at this level. Some systematic differences in effect sizes across interventions or across outcomes is to be expected, and without including fixed effects these differences could obscure the relationship between the effect sizes and study characteristics. For example, many of the largest studies were on conditional cash transfer programs, which were also often government-implemented. Without controlling for intervention, it would be unclear whether the observed negative relationship between sample size and effect size was just due to conditional cash transfer programs having small effect sizes. Effect sizes could additionally differ by outcome, even when standardized values are used, since some outcome variables may tend to have larger standard deviations than others. Further, it may be easier for some interventions to have an effect on a particular outcome, so that there is some variation in effect size by intervention-outcome. I include intervention-outcome fixed effects to abstract from any such issues in this table.

However, whether $\hat{\tau}^2_N$ or $\hat{I}^2_N$ varies systematically with any intervention-, outcome-, or intervention-outcome-level characteristics is also of interest. It is harder to analyze differences in $\hat{\tau}^2_N$ or $\hat{I}^2_N$ because there are not many intervention-outcome

and whether attrition was reported. I believe these are the most relevant variables, as the other study characteristics gathered were simply paper or result indicators, seemingly unrelated, or quite noisy (for example, the variable "number of months after intervention" was collected to capture the duration of time that had passed between the beginning of the intervention and the midline or endline data collection, however, this was unclear in many papers). The coding manual is available as an appendix for a list of other potential covariates.

35. These were: the impact of conditional cash transfers on birth at a healthcare facility; the impact of conditional cash transfers on labor hours; the impact of conditional cash transfers on pregnancy rates; the impact of conditional cash transfers on retention rates; the impact of conditional cash transfers on having a skilled attendant at delivery; the impact of financial literacy on having savings; the impact of financial literacy on having taken a loan; the impact of water treatment on diarrhea incidence; the impact of water treatment on diarrhea prevalence; and the impact of water treatment on dysentery incidence.

combinations (and hence observations of $\widehat{\tau}^2_N$ or $\widehat{I}^2_N$) that can be used in a regression, especially when using standardized values. Recall that data for 10 intervention-outcomes were unable to be standardized; this leaves 47 intervention-outcome combinations to use in a regression. Still, using k to denote intervention-outcome combinations, I can run regressions of the form:

$$\widehat{\tau}^2_N = \alpha + \beta X_k + \varepsilon_k \tag{24}$$

$$\widehat{I}^2_N = \alpha' + \beta' X'_k + \varepsilon'_k \tag{25}$$

where X_k and X'_k represent explanatory variables that vary at the intervention-outcome level.[36] To form X_k, I use the within-intervention-outcome variance of each of the explanatory variables in Table 7, in turn. The intuition is that if there is a strong relationship between study characteristics and effect size, the within-intervention-outcome variance in those characteristics might help to explain the variance in effect sizes. In place of the within-intervention-outcome variance of each of a set of regional dummy variables, however, I use the number of countries represented in an intervention-outcome, controlling for the number of studies in that intervention-outcome. This is because this measure might be easier to interpret and minimizes the number of explanatory variables. In addition, it could be that context varies immensely between countries, so that countries may be a better unit for analysis than regions.[37] X'_k is constructed as the mean value within the intervention-outcome combination of each of the explanatory variables considered in Table 7 rather than the within-intervention-outcome variance of these variables. The mean might be more appropriate for these regressions since $\widehat{I}^2_N$ captures a *proportion* of variance rather than a variance, but I show results using the within-intervention-outcome variance in an alternative specification (Table B.5 in Online Appendix B). In my preferred specification, I also winsorize one outlier for $\widehat{\tau}^2_N$.[38] Results without winsorizing this outlier are included in Table B.5 and are not very different.

Table 8 shows that the regressions of $\widehat{\tau}^2_N$ on the aforementioned explanatory variables are mostly null. It is possible that this is a result of the extra variation introduced by not including intervention or outcome fixed effects. The within-intervention-outcome variance in the sample size appears to be negatively correlated with $\widehat{\tau}^2_N$, but this is likely to be an artifact of CCTs having the greatest variance in sample size and also having relatively low $\widehat{\tau}^2_N$. $\widehat{I}^2_N$ is positively associated with the mean sample size within an intervention-outcome combination. This is not a

36. $\widehat{\tau}^2_N$ and $\widehat{I}^2_N$ naturally vary at the intervention-outcome level; they could equally well be subscripted as $\widehat{\tau}^2_{kN}$ and $\widehat{I}^2_{kN}$.

37. Country dummies were not included in Table 7 because they would have been likely to result in overfitting.

38. One value of $\widehat{\tau}^2_N$ is 6.6 standard deviations away from the mean and several times higher than the next largest value, so it may make sense to treat as an outlier. This $\widehat{\tau}^2_N$ was estimated for the impact of rural electrification programs on study time and seems to be a result of studies finding impacts ranging from a few minutes to several hours per week.

TABLE 7. 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)	
Observations	528	597	611	528	521
R^2	0.19	0.22	0.21	0.21	0.19

Notes: 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. This table uses those intervention-outcomes covered by at least 2 papers; readers will recall the maximum number of observations for this data set was 612, before dropping the one outlier. 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. Intervention-outcome fixed effects are included, with standard errors clustered by intervention-outcome.

surprise, because increases in sample size reduce the sampling variance and so should mechanically reduce $\hat{I}^2_N$, independent of the relationship of sample size to $\hat{\tau}^2_N$. In the alternative specification in Table B.5 that uses the within-intervention-outcome variance rather than the mean as an explanatory variable, the academic/NGO-implemented variable is also significantly associated with $\hat{I}^2_N$, but this could be due to those intervention-outcomes with high variance in implementation containing more government-implemented programs and government-implemented programs tending to have larger sample sizes, driving down the sampling variance and hence driving up $\hat{I}^2_N$.

The explanatory variables used in Table 8 are not the only ones that might have a theoretical reason to be associated with $\hat{\tau}^2_N$. A stronger relationship might hold between $\hat{\tau}^2_N$ and how direct an impact an intervention is likely to have. Those intervention-outcome combinations for which the interventions act more directly on the targeted outcomes may be expected to have smaller $\hat{\tau}^2_N$. This hypothesis has frequently been made in the literature on "theories of change" or "causal chains" e.g., Williams (2018). However, it is difficult to operationalize this intuition. I focus on two examples for which I think there is theoretical reason to believe the effect of the intervention on certain outcomes is particularly direct: the effect of health interventions and the effect

TABLE 8. Regression of $\hat{\tau}^2_N$ and $\hat{I}^2_N$ on study characteristics.

	$\hat{\tau}^2_N$					
	(1)	(2)	(3)	(4)	(5)	(6)
Var(Sample Size)	-0.045**					-0.026
	(0.02)					(0.06)
Var(Government-implemented)		0.118				0.651
		(0.40)				(0.80)
Var(Academic/NGO-implemented)			0.019			-0.685
			(0.36)			(0.44)
Var(RCT)				-0.268		-0.144
				(0.40)		(0.58)
Number of Countries					-0.033	-0.019
					(0.03)	(0.04)
Number of Studies					0.006	0.001
					(0.01)	(0.02)
Observations	41	47	47	47	47	41
R^2	0.01	0.00	0.00	0.01	0.11	0.12
	$\hat{I}^2_N$					
	(7)	(8)	(9)	(10)	(11)	(12)
Mean(Sample Size)	0.094*					0.139**
	(0.05)					(0.06)
Mean(Government-implemented)		0.026				-0.154
		(0.06)				(0.11)
Mean(Academic/NGO-implemented)			-0.056			-0.057
			(0.06)			(0.14)
Mean(RCT)				-0.066		-0.073
				(0.09)		(0.14)
Number of Countries					-0.008	-0.017
					(0.01)	(0.02)
Number of Studies					0.004	0.008
					(0.01)	(0.01)
Observations	41	47	47	47	47	41
R^2	0.02	0.00	0.02	0.01	0.00	0.06

Notes: This table shows the results of regressions of $\hat{\tau}^2_N$ and $\hat{I}^2_N$ on intervention-outcome-level summary statistics of the study characteristics considered in Table 7 (i.e., estimating $\hat{\tau}^2_N = \alpha + \beta X_k + \varepsilon_k$ and $\hat{I}^2_N = \alpha' + \beta' X'_k + \varepsilon'_k$ where X_k and X'_k represent intervention-outcome-level summary statistics such as the variance of the sample size of studies within an intervention-outcome). One outlier value of $\hat{\tau}^2_N$ 6.6 standard deviations away from the mean is winsorized, as described in the text. Robust standard errors are used.

of interventions that provide an economic incentive that is conditional. It is frequently hypothesized that results from social science interventions vary more than results for interventions that produce effects through biological channels. From an economic standpoint, conditional programs that have a direct causal mechanism through which they are posited to work could also have more generalizable results.

To test these hypotheses, I regress $\hat{\tau}^2_N$ and $\hat{I}^2_N$ on dummy variables indicating

whether the intervention in the intervention-outcome combination in question is a health intervention or a conditional intervention. These regressions take the same form as Equations 24 and 25, but now k is used to denote interventions rather than intervention-outcomes and X_k and X'_k each indicate whether the intervention is a health or conditional intervention, in turn.

I label as health interventions deworming drugs, micronutrient supplementation programs and bed nets programs. HIV/AIDS education programs might also be thought of as health interventions, though they are based on behavior change rather than on direct provision of drugs or supplements, and school meals programs also are somewhat health-related. In an alternative specification, I include these latter two interventions as health interventions. I classify conditional cash transfer programs (which generally provide benefits conditional on enrollment in school) and performance pay programs (which provide benefits conditional on test scores) as conditional programs. One may also consider scholarships programs to be implicitly conditional given that one needs to continue to attend school in order to receive the scholarship. I include scholarships as a conditional intervention in an alternative specification.

The main regression results are reported in Table 9, while Appendix Table B.6 provides results without winsorizing one value of $\hat{\tau}^2_N$ and Appendix Table B.7 provides results for the regressions using the alternative definitions of health and conditional interventions. There is some suggestive evidence that health interventions and conditional economic interventions have lower $\hat{\tau}^2_N$. However, these results are sensitive to whether one winsorizes an extreme outlier for $\hat{\tau}^2_N$ and whether the alternative definitions are used. The point estimates all have the expected sign: health interventions and conditional economic interventions have smaller $\hat{\tau}^2_N$. It remains possible that outcomes with lower standardized $\hat{\tau}^2_N$ are simply overrepresented in the outcomes studied by these interventions. No significant relationship is observed with $\hat{I}^2_N$.[39]

These tables illustrate that it is not easy to make quick judgments about which types of interventions generalize. Health interventions have long been suspected to be distinctly better at obtaining generalizable results than interventions that act through social or behavioral pathways. I find some evidence of this, but the fact that the relationship is not stronger suggests that the story is not so straightforward. One possible explanation is that the results of health interventions can depend greatly on the baseline prevalence of the disease they were intended to treat, and these regressions do not control for that. This motivates the next stage of analysis: modeling within-intervention-outcome variation.

4.2.2. Within Intervention-Outcomes. While seeking to explain heterogeneity across intervention-outcomes has the advantage of enabling a larger sample of studies to be

39. It can be observed that the sign of the relationship flips for conditional programs. This could be a function of the CCTs captured by this variable tending to have large sample sizes, which would increase $\hat{I}^2_N$.

TABLE 9. Regression of $\hat{\tau}^2_N$ and $\hat{I}^2_N$ on intervention characteristics.

	$\hat{\tau}^2_N$			$\hat{I}^2_N$		
	(1)	(2)	(3)	(4)	(5)	(6)
Health	-0.114		-0.210*	-0.074		-0.086
	(0.09)		(0.12)	(0.05)		(0.05)
Conditional		-0.128**	-0.262**		0.023	-0.032
		(0.05)	(0.12)		(0.05)	(0.05)
Observations	47	47	47	47	47	47
R^2	0.04	0.03	0.13	0.04	0.00	0.05

Notes: This table shows the results of regressions of $\hat{\tau}^2_N$ and $\hat{I}^2_N$ on intervention-level characteristics (i.e., estimating $\hat{\tau}^2_N = \alpha + \beta X_k + \varepsilon_k$ and $\hat{I}^2_N = \alpha' + \beta' X'_k + \varepsilon'_k$ where X_k and X'_k now represent the intervention-level characteristics of whether the intervention was a health intervention and whether it provided economic incentives that were conditional on certain actions). As before, one value of $\hat{\tau}^2_N$ 6.6 standard deviations away from the mean is winsorized. Robust standard errors are used.

used, more variation might be explained if I modeled heterogeneity within particular intervention-outcome combinations. To this end, I focus on those intervention-outcome combinations covered by over 10 studies. I exclude micronutrient programs so as to focus on those interventions more often studied by economists. To explain heterogeneity in treatment effects across studies within an intervention-outcome, I leverage both the potential explanatory variables that are shared in common across all the intervention-outcome combinations, used in the previous regressions, and the variables that were coded that are intervention-specific. Excluding micronutrients, only CCTs, UCTs, and deworming programs have over 10 studies on a particular outcome in my data. Table 10 lists the intervention-specific variables that were coded for each of these interventions.[40]

Some of the intervention-specific variables relate to the sample (e.g., age, gender). Variables relating to the sample often varied within a study and different values were reported for different subgroups. To generate a study-level aggregate value, the same process was followed as was used to create a single treatment effect per intervention-outcome-paper, creating a weighted mean. If a paper reported aggregate values alongside results for subgroups, the aggregate value was used, else the smallest set of non-overlapping subgroups were aggregated. For example, if results were reported separately for girls and boys and also for three different age groups, the results for girls and boys would be aggregated.

With these variables, I can estimate decreases in τ^2 and I^2 were a mixed model to be used. With many possible explanatory variables and a small number of observations, I must select among the explanatory variables. I choose the single explanatory variable

40. In addition to these variables, I also consider the possibility that there is an interaction between the drug provided and the dosage, since different drugs have different strengths and are typically given in different amounts.

TABLE 10. List of intervention-specific variables.

CCTs:
Minimum transfer per child conditional on meeting education requirements
Maximum transfer per child conditional on meeting education requirements
Minimum transfer per household conditional on meeting education requirements
Maximum transfer per household conditional on meeting education requirements
Min transfer per household conditional on meeting non-education-related requirements
Max transfer per household conditional on meeting non-education-related requirements
Whether program eligibility was restricted to poor households
Whether enrollment at school was a condition
Whether attendance at school was a condition
What the threshold attendance level was for those conditional on school attendance
Whether there were any health-related conditions, such as health checks
Baseline enrollment rates
Whether the sample comprised only those enrolled at baseline, not enrolled, or a mix
Whether the study was done in a rural and/or urban setting
Results for other programs in the same region
The age range of the sample under consideration
The gender of the sample under consideration
UCTs:
The minimum transfer amount per child
The maximum transfer amount per child
The minimum transfer amount per household
The maximum transfer amount per household
Whether program eligibility was restricted to poor households
Baseline enrollment rates
Whether the sample comprised only those enrolled at baseline, not enrolled, or a mix
Whether the study was done in a rural and/or urban setting
Results for other programs in the same region
The age range of the sample under consideration
The gender of the sample under consideration
Deworming:
Indicators for albendazole, mebendazole, levamisole, pyrantel pamoate, or multiple drugs
How many rounds of treatment there were
How many months elapsed between each round
The dosage of each drug provided in one round
The baseline prevalence of each of *Ascaris lumbricoides*, *Trichuris trichiura*, and hookworm

Notes In some cases, only endline enrollment or prevalence rates are reported. The baseline rates variables are therefore constructed by using baseline rates for both the treatment and control group where they are available, followed by 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. Regions include Latin America, Africa, the Middle East and North Africa, East Asia, and South Asia, following the classification of World Bank (2015).

which maximizes the R^2 when running an OLS regression of the treatment effect Y_i on the explanatory variable X_i ($Y_i = \alpha + \beta X_i + \varepsilon_i$) run separately within each intervention-outcome. The "residual" $\hat{\tau}^2$, $\hat{\tau}^2_R$, is then calculated using the mixed model described by Equation 15, with the selected explanatory variable as its X_i.

Results are presented in Table 11. On average, $\hat{\tau}^2_R$ is reduced from $\hat{\tau}^2$ by about

20%. The median is a bit lower at 10%, as there are several intervention-outcomes for which $\hat{\tau}^2$ does not appreciably decrease, and $\hat{\tau}^2_R$ and $\hat{I}^2_R$ actually minutely increase for two intervention-outcome combinations, reflecting simulation noise. The intervention-outcome combinations for which $\hat{\tau}^2_N$ decreased the most were the impact of deworming on weight-for-height (73%), the impact of deworming on weight (68%), and the impact of deworming on height-for-age (25%). It should be noted that while I restricted attention to those intervention-outcome combinations with over 10 studies, many of the papers failed to report all the explanatory variables, reducing the effective number of observations. There is thus a risk that some of the largest decreases are the result of overfitting.

Given the number of studies within an intervention-outcome combination, it is infeasible to build models with more explanatory variables. Further gains may, however, be possible by leveraging micro-data when they are available. To provide support for this intuition, I turn to consider an intervention-outcome combination covered by a particularly large number of studies: the effect of conditional cash transfers on enrollment rates. While up to this point the paper has used data that either were originally reported as an aggregate point estimate or data from combining the minimum number of non-overlapping subgroups, here I turn to consider the maximal set of non-overlapping subgroups to increase the sample size and run OLS regressions of the unstandardized treatment effect on these sample characteristics.[41] If Y_{is} is the estimated effect of a conditional cash transfer program on enrollment rates in subgroup s of study i, these regressions are of the form: $Y_{is} = \alpha + \beta X_{is} + \varepsilon_{is}$. These subgroups never overlap, so no results are double-counted, but results can be correlated across subgroups within a study, so I cluster standard errors by study. Since most variables describing a paper (such as whether it was an RCT) do not vary within the paper, I consider only those variables that describe sample characteristics as explanatory variables in these regressions.

Results are presented in Table 12. The baseline enrollment rates show the strongest relationship to the treatment effect, as reflected in the R^2 of these regressions and their significance levels. It seems easier for there to be large treatment effects when the baseline level of the outcome variable is 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 having a 0% or 100% enrollment rate at baseline, respectively, in addition to being represented by two dummy variables. CCT programs seem to have larger effects on enrollment rates for those not enrolled at baseline, beyond the linear trend (Column 2). Studies done in urban areas tend to find smaller treatment effects than studies done in rural or mixed urban/rural areas. There is no significant difference in treatment effects by gender or age.[42] Finally, for each observation I calculate the mean treatment effect in the same region, excluding results from the program in question.

41. For example, if I have results for girls and boys reported separately, as well as for three different age groups, I will now use the results for the three different age groups.

42. Shown here: minimum sample age. Results for the maximum or mean age variables available upon request.

Treatment effects do appear correlated across different programs in the same region.

If data that are disaggregated even just to the subgroup level can obtain a much improved fit, it would suggest that models leveraging micro-data would yield even better results.[43]

43. I do not run a mixed model using the significant characteristics as explanatory variables because doing so would artificially increase the estimated τ^2 for the intervention-outcome. For example, splitting the papers' samples into results by age group would generally serve to increase τ^2 relative to using the aggregate result. Even if that estimated τ^2 could then be reduced by the mixed model, it is not clear what the implication would be for the perhaps more standard scenario in which one wants to compare aggregate point estimates across papers.

TABLE 11. Residual heterogeneity measures by intervention-outcome.

Intervention-Outcome	Explanatory Variable	R^2	$\hat{\tau}^2$	$\hat{\tau}^2_R$	$\frac{\hat{\tau}^2-\hat{\tau}^2_R}{\hat{\tau}^2}$	$\hat{I}^2$	$\hat{I}^2_R$	$\frac{\hat{I}^2-\hat{I}^2_R}{\hat{I}^2}$	N
CCTs-Attendance rate	Baseline enrollment rate	0.43	0.0031	0.0029	0.08	0.879	0.856	0.03	8
CCTs-Enrollment rate	Min household non-educ. Transfer	0.28	0.0010	0.0008	0.20	0.961	0.952	0.01	36
CCTs-Labor force particip.	Conditional on health check	0.38	0.0012	0.0013	-0.07	0.938	0.944	-0.01	10
UCTs-Enrollment rate	Sample minimum age	0.34	0.0006	0.0006	0.04	0.845	0.848	0.00	10
Deworming-Height	Mebendazole dosage	0.32	0.2201	0.2097	0.05	0.942	0.940	0.00	13
Deworming-Height-for-age	Mix of drugs	0.32	0.0497	0.0373	0.25	0.989	0.986	0.00	13
Deworming-Hemoglobin	Baseline prevalence *T. Trichiura*	0.36	0.0077	0.0083	-0.07	0.643	0.656	-0.02	11
Deworming-Weight	Baseline prevalence hookworm	0.73	0.3596	0.1153	0.68	0.995	0.984	0.01	9
Deworming-Weight-for-age	Baseline prevalence *T. Trichiura*	0.39	0.0114	0.0101	0.11	0.966	0.960	0.01	8
Deworming-Weight-for-height	Baseline prevalence hookworm	0.92	0.0191	0.0052	0.73	0.911	0.604	0.34	5

Notes: This table shows residual heterogeneity measures after fitting a mixed model to each of several intervention-outcome combinations with a particularly large number of studies. Each mixed model used the single explanatory variable with the highest R^2 in an OLS regression of treatment effects on each potential explanatory variable in Table 10. The explanatory variable selected is reported in this table, along with the R^2 and number of observations available for this regression. One aggregate result per intervention-outcome-study is used.

TABLE 12. Regression of projects' results on characteristics (CCTs on enrollment rates).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Enrollment Rates	-0.205***	-0.102***								-0.090***	-0.081**
	(0.05)	(0.03)								(0.03)	(0.03)
Enrolled at Baseline		0.001									-0.002
		(0.02)									(0.02)
Not Enrolled at		0.195***								0.199***	0.189***
Baseline		(0.03)								(0.02)	(0.04)
Number of			-0.008								0.003
Observations (100,000s)			(0.00)								(0.00)
Rural				0.038**						0.013	0.032
				(0.02)						(0.01)	(0.02)
Urban					-0.049***					-0.018	-0.017
					(0.01)					(0.01)	(0.02)
Girls						0.001					0.014
						(0.01)					(0.01)
Boys							-0.020				-0.004
							(0.02)				(0.02)
Minimum Sample Age								0.001			0.002
								(0.00)			(0.00)
Mean Regional Result									1.000**		0.263
									(0.47)		(0.36)
Observations	249	249	145	270	270	270	270	244	270	249	119
R^2	0.32	0.44	0.00	0.05	0.03	0.00	0.01	0.00	0.02	0.45	0.50

Notes: Each column regresses the impact of conditional cash transfer programs on enrollment rates (i.e., the subgroup-level point estimates Y_{is}) on different explanatory variables. Multiple results for different subgroups may be reported for the same paper. The data on which this table is based include 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. "Enrolled at Baseline" is a dummy variable indicating whether the entire sample on which a result was reported was enrolled in school at baseline; "Not Enrolled at Baseline" is a dummy variable indicating whether the entire sample was not enrolled in school at baseline. These correspond to 100% and 0% enrollment rates for the sample under consideration, but it makes sense to consider them separately due to selection issues.

5. Discussion

Why should we care about the dispersion of results across studies? Information on the context, intervention, implementation and study quality could decrease the amount of unexplained heterogeneity and so improve inference. However, the information provided in academic papers is naturally limited. I find that sampling variance accounts for only 6% of the total variance in estimated treatment effects for the median intervention-outcome combination in my data. For 10 intervention-outcome combinations in my data with a large number of studies, I find about 20% of the remaining variance could be explained using a single best-fitting explanatory variable. However, this statistic obscures a lot of heterogeneity, with the median decrease being about 10%. These results, though perhaps better than many might expect, emphasize the importance of sharing micro-data to build even better models and treating the point estimates reported in papers as merely a starting point.

A few limitations should be discussed. I classify different programs as the "same" intervention despite minor differences between them. This is because these programs differ in too many idiosyncratic ways to be able to usefully categorize them into finer groups. While describing multiple distinct programs as being part of the "same" intervention may be a common practice, it is important to remember that some of the observed variation could be due to differences in the programs themselves.

One may also be concerned that results are driven by those interventions with the greatest number of studies in the data set: micronutrients programs, conditional cash transfers, and deworming programs. For results that are presented disaggregated by intervention-outcome, such as the results in Table 4, this is not a concern. For the regressions across intervention-outcome combinations reported in Table 7, however, one might still wonder if results were driven by these interventions. It is also possible that outcomes within the same intervention may be correlated. As discussed, I cannot combine outcomes within an intervention, as that would make it harder to determine the source of the observed heterogeneity.

More heterogeneity in treatment effects might be modeled using micro-data; data taken from the results reported in academic papers are not as rich, providing both fewer observations and fewer covariates. However, despite shifting norms, micro-data are still rarely available,[44] so the approach outlined in this paper may still frequently be useful. Further, if more authors were to make their study's micro-data available, that would not in and of itself be sufficient for building more complicated models to explain heterogeneity, as typically there is little overlap in covariates collected across different studies. To remedy this, more collaboration among researchers at an earlier stage would be helpful.

In light of the observed variation in studies' results, how policymakers combine information from different studies is a fruitful area for further research. Coville and Vivalt (2019), for example, find some evidence that policymakers exhibit "variance

44. Less than 10% of the studies in AidGrade's database made micro-data available.

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. 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 hence underscores the importance of further research to determine how policymakers interpret the information they are given and how to best present information to enable optimal decision-making.

6. Conclusion

How much impact evaluation results generalize to other settings is an important question. Before now, no data set existed with many different types of interventions, with all data 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 results presented here, it begins to be possible 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.

I 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, as some policy decisions might have greater room for error than others. However, I suggested a way of thinking about the problem based on the relationship between these measures and one's ability to draw inferences about results in another setting and provided estimates for many intervention-outcome combinations.

I 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 might be expected 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 significantly different effect sizes than quasi-experimental studies.

I then sought to explain heterogeneity within several intervention-outcome combinations covered by a large number of studies. Heterogeneity measures greatly improved for some intervention-outcome combinations, but not for others. I also explored variation across different subgroups for one particular intervention-outcome combination, which afforded a larger sample size. Taken together, the results suggest that careful modeling could help substantially and that there are likely to be large gains

in using even more disaggregated micro-data.

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 et al. 2012), one solution would be to ensure one's impact evaluation varied some of the contextual variables that one 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. Framing these efforts as increasing our understanding 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

Allcott, Hunt (2015). "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics*, 130(3), 1117–1165.

Bandiera, Oriana, Greg Fisher, Andrea Prat, and Erina Ytsma (2016). "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments." *Working paper*.

Banerjee, Abhijit V. and Esther Duflo (2009). "The Experimental Approach to Development Economics." *Annual Review of Economics*, 1(1), 151–178.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur (2018). "Experimental Evidence on Scaling up Education Reforms in Kenya." *Journal of Public Economics*, 168, 1–20.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein (2009). *Introduction to Meta-Analysis*. Wiley John + Sons, URL https://www.ebook.de/de/product/8016062/michael_borenstein_larry_v_hedges_julian_p_t_higgins_hannah_r_rothstein_introduction_to_meta_analysis.html.

Briggs, Derek C. and Mark Wilson (2007). "Generalizability in Item Response Modeling." *Journal of Educational Measurement*, 44(2), 131–155.

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.

Combs, James G., David J. Ketchen Jr, T. Russell Crook, and Philip L. Roth (2009). "Assessing Cumulative Evidence within 'Macro' Research: Why Meta-Analysis Should be Preferred Over Vote Counting." *Journal of Management Studies*, 48(1), 178–197.

Coville, Aidan and Eva Vivalt (2019). "How do Policymakers Update?" *Working paper*.

Deaton, Angus (2010). "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature*, 48(2), 424–455.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii (2019). "From Local to Global: External Validity in a Fertility Natural Experiment." *Journal of Business & Economic Statistics*, pp. 1–27.

Efron, Bradley and Carl Morris (1975). "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association*, 70(350), 311–319.

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

Gelman, Andrew (2006). "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis*, 1(3), 515–534.

Gelman, Andrew and John Carlin (2014). "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science*, 9(6), 641–651.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, and Aki Vehtari (2013). *Bayesian Data Analysis*. 3 ed., Taylor & Francis Ltd, URL https://www.ebook.de/de/product/15022612/andrew_department_of_statistics_columbia_university_new_york_usa_gelman_john_b_the_royal_children_s_hospital_parkville_victoria_australia_carlin_hal_s_university_of_california_irvine_usa_stern_david_b_duke_university_durham_north_carolina_usa_dunson_aki_aalto_university_finland_vehtari_bayesian_data_analysis.html.

Gelman, Andrew and Francis Tuerlinckx (2000). "Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures." *Computational Statistics*, 15(3), 373–390.

Hedges, Larry V. and Ingram Olkin (1980). "Vote-counting Methods in Research Synthesis." *Psychological Bulletin*, 88(2), 359–369.

Higgins, JPT and S Green (2011). "Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0." URL http://handbook.cochrane.org.

Higgins, Julian P. T. and Simon G. Thompson (2002). "Quantifying Heterogeneity in a Meta-analysis." *Statistics in Medicine*, 21(11), 1539–1558.

Jadad, A R, R A Moore, D Carroll, C Jenkinson, D J Reynolds, D J Gavaghan, and H J McQuay (1996). "Assessing the quality of reports of randomized clinical trials: is blinding necessary?" *Controlled Clinical Trials*, 17, 1–12.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van 't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka, and Brian A. Nosek (2014). "Investigating Variation in Replicability." *Social Psychology*, 45(3), 142–152.

Koricheva, Julia and Jessica Gurevitch (2013). *Handbook of Meta-analysis in Ecology and Evolution*, chap. Place of Meta-analysis among other Methods of Research Synthesis, pp. 3–13. Princeton University Press.

Kowalski, Amanda (2016). "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments." Tech. rep.

Meager, Rachael (2019). "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics*, 11(1), 57–91.

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*.

Rubin, Donald B. (1981). "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics*, 6(4), 377.

Shavelson, Richard J. and Noreen M. Webb (1991). *Generalizability Theory: A Primer*. Sage Publications, URL https://www.ebook.de/de/product/3803890/richard_j_shavelson_noreen_m_webb_generalizability_theory_a_primer.html.

Stein, Charles (1956). "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 197–206, University of

California Press, Berkeley, Calif., URL https://projecteuclid.org/euclid.bsmsp/1200501656.

Vivalt, Eva (2019). "Specification Searching and Significance Inflation Across Time, Methods and Disciplines." *Oxford Bulletin of Economics and Statistics*, 81(4), 797–816.

World Bank (2015). "Country and Lending Groups." http://data.worldbank.org/about/countryand-lending-groups.