How Much Can We Generalize From Impact Evaluations?

Eva Vivalt

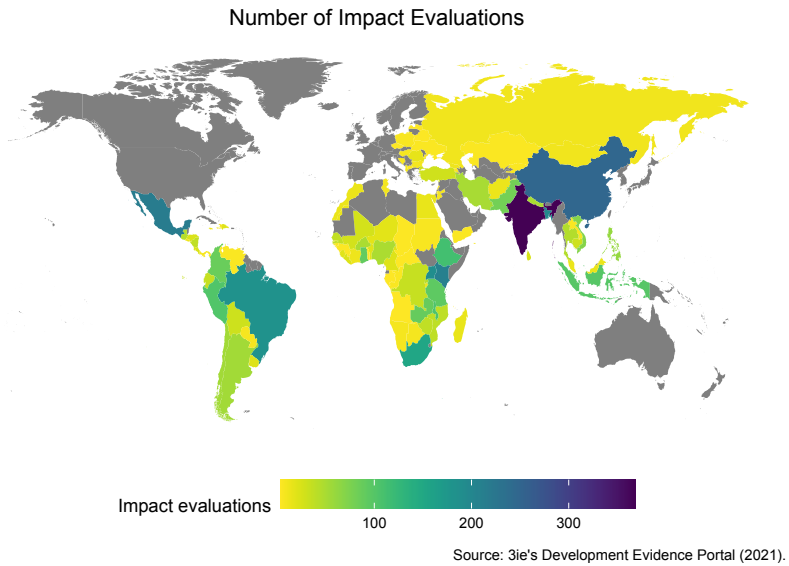
University of Toronto

Research Questions

- How much can we generalize?
- Where is the variation coming from?
  - Implementation/context differences?
  - Sampling error?
  - Specification searching/publication bias?

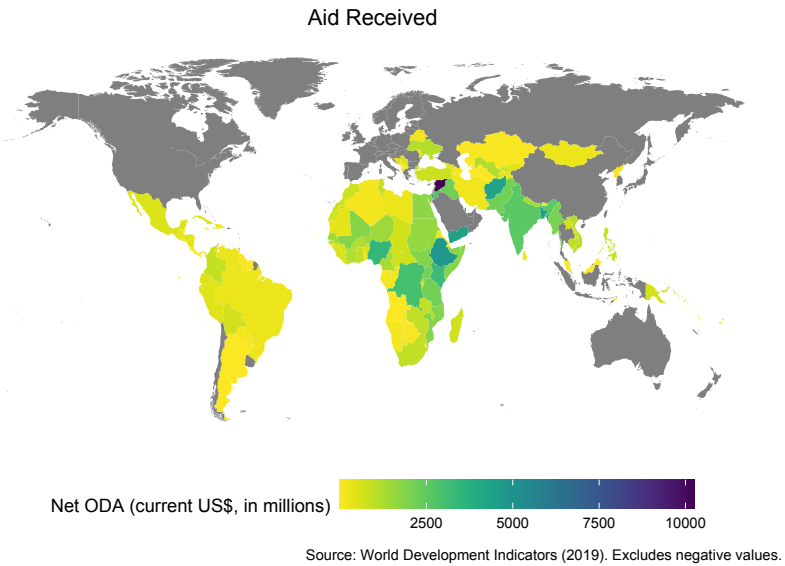
Motivation

- Impact evaluations used to inform future work
- Results vary
- If we don't know why, don't know what will happen when implementing that project in a different context



Motivation

- Impact evaluations used to inform future work
- Results vary
- If we don't know why, don't know what will happen when implementing that project in a different context



Motivation

Concerns about external validity:

- Example of same place, different effects (Bold *et al.*, 2018)
- Site selection bias (Allcott, 2015)
- Specific contexts like conditional cash transfers (CCTs)
- General critiques:
  - Economics (Deaton, 2011; Sandefur and Pritchett, 2013)
  - Other social sciences, health (Campbell and Stanley, 1963; CONSORT, 2010)

But *how much* can we generalize in practice?

New Data

- AidGrade's data set of impact evaluation results, gathered in the course of meta-analyses
- 635 studies on 20 types of interventions

Table: List of Development Programs Covered

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

Strategy

- Begin by discussing common measures of heterogeneity from the meta-analysis literature
- Relate these measures to generalizability
- Generate statistics for each intervention-outcome combination

Takeaways

1. Results vary more than one might expect:
   - One would guess the correct sign 61% of the time
   - The median ratio of the root-MSE to the meta-analysis mean is 2.49
2. Not much of the variance is due to sampling variance (6%)
3. Modest improvement using a mixed model ($\sim$ 20% on average, 10% median across intervention-outcomes)
4. Larger projects do worse
5. Academic/NGO-implemented projects do better than government-implemented projects
6. Some types of interventions do better

Random-Effects Meta-Analysis Model

$$Y_i = \theta_i + u_i$$
$$u_i \sim N(0, \sigma_i^2)$$
$$\theta_i \sim N(\mu, \tau^2)$$

Y_i is the estimate of the effect in study i

θ_i is the true effect in study i

u_i is the error, normally distributed with some sampling variance σ_i^2

μ is the grand mean

τ^2 is the inter-study variance

Mixed Model

$$Y_i = \theta_i + u_i$$

$$\theta_i = \alpha + X_i\beta + e_i$$

$$u_i \sim N(0, \sigma_i^2)$$

$$e_i \sim N(0, \tau_R^2)$$

Y_i is the estimate of the effect in study i

θ_i is the true effect in study i, and it has some component that can be explained ($X_i\beta$) and some component that cannot (e_i)

u_i is the error, normally distributed with some sampling variance σ_i^2

τ_R^2 is the (residual) inter-study variance after accounting for $X_i\beta$

Measuring Generalizability

- How should we define generalizability?
- How can we relate it to heterogeneity measures?

Classical Measures of Heterogeneity

Two classes of measures:

- Variation
  - Variance in effect sizes Y_i
  - True inter-study variance τ^2
  - Coefficient of variation: standard deviation/mean or τ/μ
- Proportion of variation
  - I^2: $\frac{\tau^2}{\sigma^2+\tau^2}$, where τ^2 is the true variance of effect sizes and σ^2 captures sampling error.

Classical Measures of Heterogeneity

Two classes of measures:

- Variation
  - Variance in effect sizes Y_i
  - True inter-study variance τ^2
  - Coefficient of variation: standard deviation/mean or τ/μ
- Proportion of variation
  - I^2: $\frac{\tau^2}{\sigma^2+\tau^2}$, where τ^2 is the true variance of effect sizes and σ^2 captures sampling error.

Can also create similar statistics after taking explanatory variables into consideration (e.g. "residual" τ^2, τ^2_R)

Heterogeneity Measures

Table: 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		✓	✓

Relating Generalizability to Heterogeneity Measures

- Inspiration: Gelman and Carlin (2014) and Gelman and Tuerlinckx (2000)'s Type S and Type M errors
  - Type S error: error in sign
  - Type M error: error in magnitude
- They consider whether a result is likely to replicate
- This can be thought of as "generalizability to the same context"
- Straightforward to extend to "generalizability to different contexts"

Relating Generalizability to Heterogeneity Measures

- The probability that an inference about an impact in another setting will have the right sign or be a certain magnitude bigger or smaller than the true value depends on the parameters in the Bayesian model: τ^2, μ, σ_i^2 (or I^2)
- So we can estimate values for these variables and then talk of inference errors of sign and magnitude

Relating Generalizability to Heterogeneity Measures

- Inference errors of sign and magnitude are highly policy-relevant
- They can be compared across intervention-outcomes
- The likely sign and magnitude of an impact are not the *only* policy-relevant questions we may be interested in. Same approach can be applied to other questions

Estimating a Random-Effects Model

Recall the basic model:

$$Y_i = \theta_i + u_i$$
$$u_i \sim N(0, \sigma_i^2)$$
$$\theta_i \sim N(\mu, \tau^2)$$

I'll estimate μ, τ^2 and θ_i using Bayesian hierarchical models

Prior for θ_i

Assume between-study normality where μ and τ are unknown hyperparameters:

$$\theta_i \sim N(\mu, \tau^2) \tag{1}$$

Likelihood for θ_i

Assume data are normally distributed:

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

Posterior for θ_i

$$\theta_i|\mu,\tau,Y \sim N(\hat{\theta}_i, V_i) \tag{3}$$

where

$$\hat{\theta}_i = \frac{\frac{Y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}, \; V_i = \frac{1}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}$$

Hierarchical Bayesian Model

Priors for $\mu|\tau$ and τ: uniformly distributed.

Update based on the data.

More

Computation:

1. $\tau|Y$
2. $\mu|\tau, Y$
3. $\theta|\mu, \tau, Y$

Mixed Model

Similar.

For random effects, used:

$$P(\theta, \mu, \tau|Y) = P(\theta|\mu, \tau, Y)P(\mu|\tau, Y)P(\tau|Y)$$

For mixed model:

$$P(e, \beta, \tau|Y) = P(e|\beta, \tau, Y)P(\beta|\tau, Y)P(\tau|Y)$$

Computation:

1. $\tau|Y$
2. $\beta|\tau, Y$
3. $e|\beta, \tau, Y$

Data

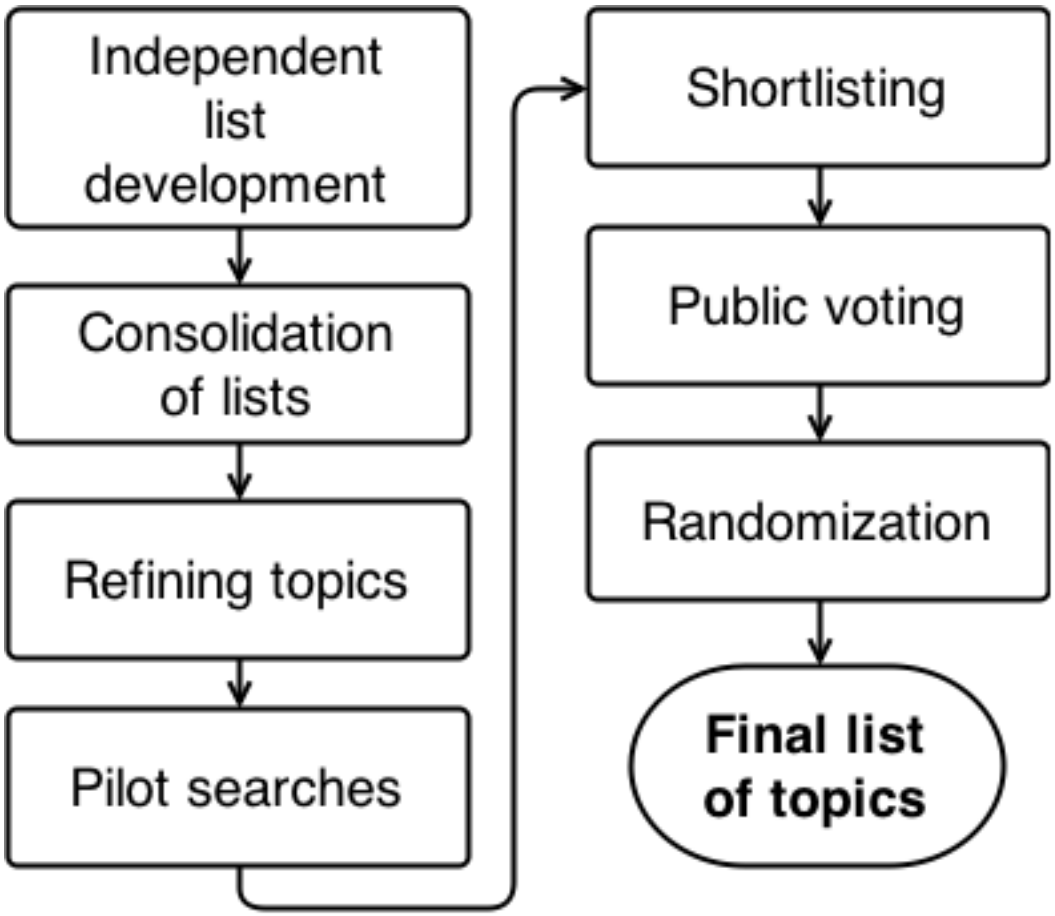
- 20 interventions
- Any impact evaluation attempting to measure counterfactual is included - experimental and quasi-experimental
- Published papers and working papers
- 85 base fields were coded for each paper
- Additional topic-specific fields to capture heterogeneity in programs and samples (frequently sparse)
- Followed Cochrane
- Double-entry coding for everything

Process

- Selection of interventions
- Search
- Screening
- Data extraction

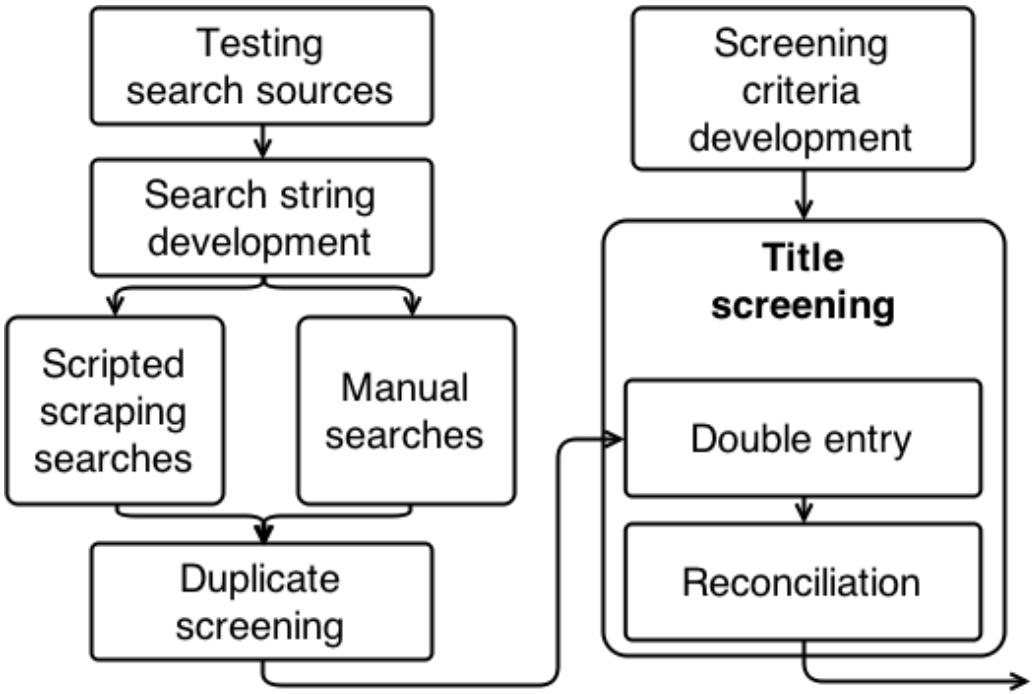
Process Diagram

Figure: Topic Selection



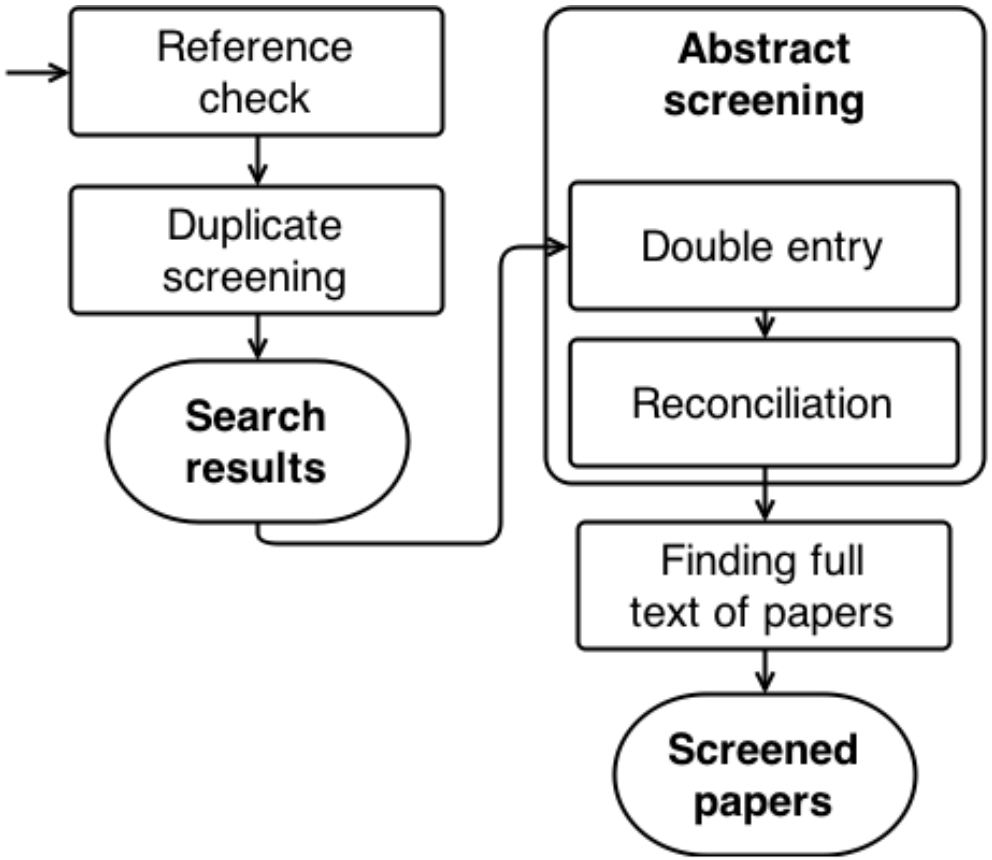
Process Diagram

Figure: Search and Screening, Part 1



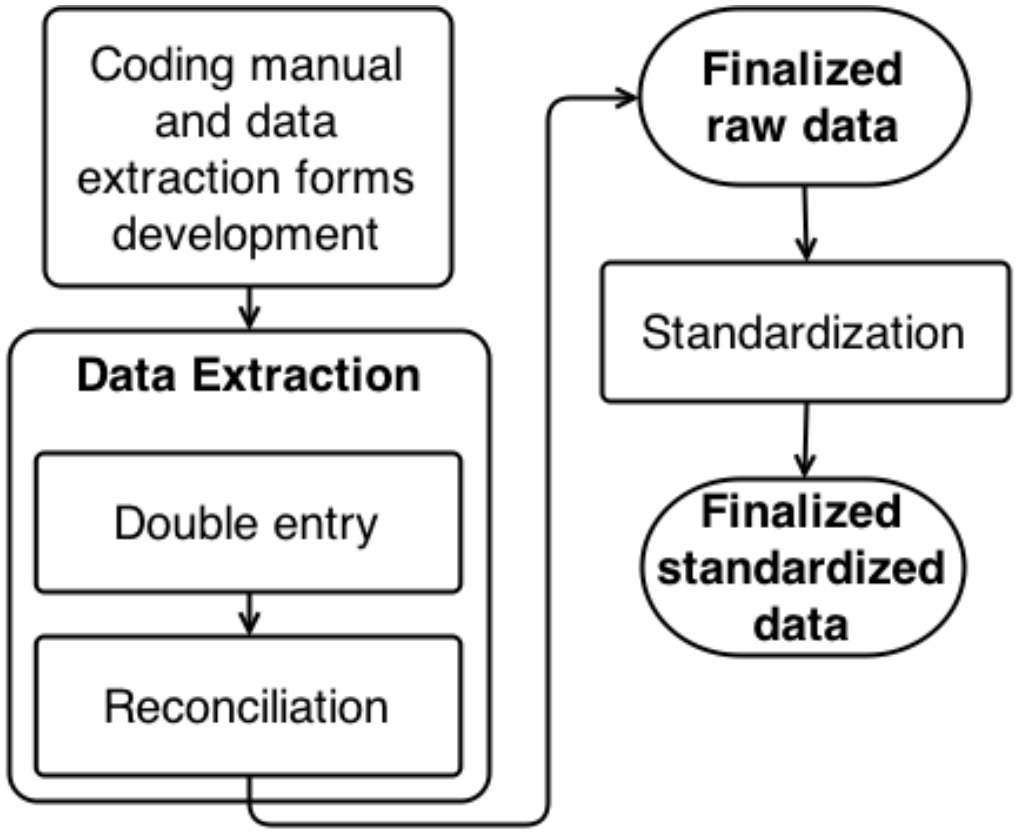
Process Diagram

Figure: Search and Screening, Part 2



Process Diagram

Figure: Data Extraction



Data

- Need to standardize effect sizes:

$$SMD = \frac{\mu_1 - \mu_2}{\sigma_p}$$

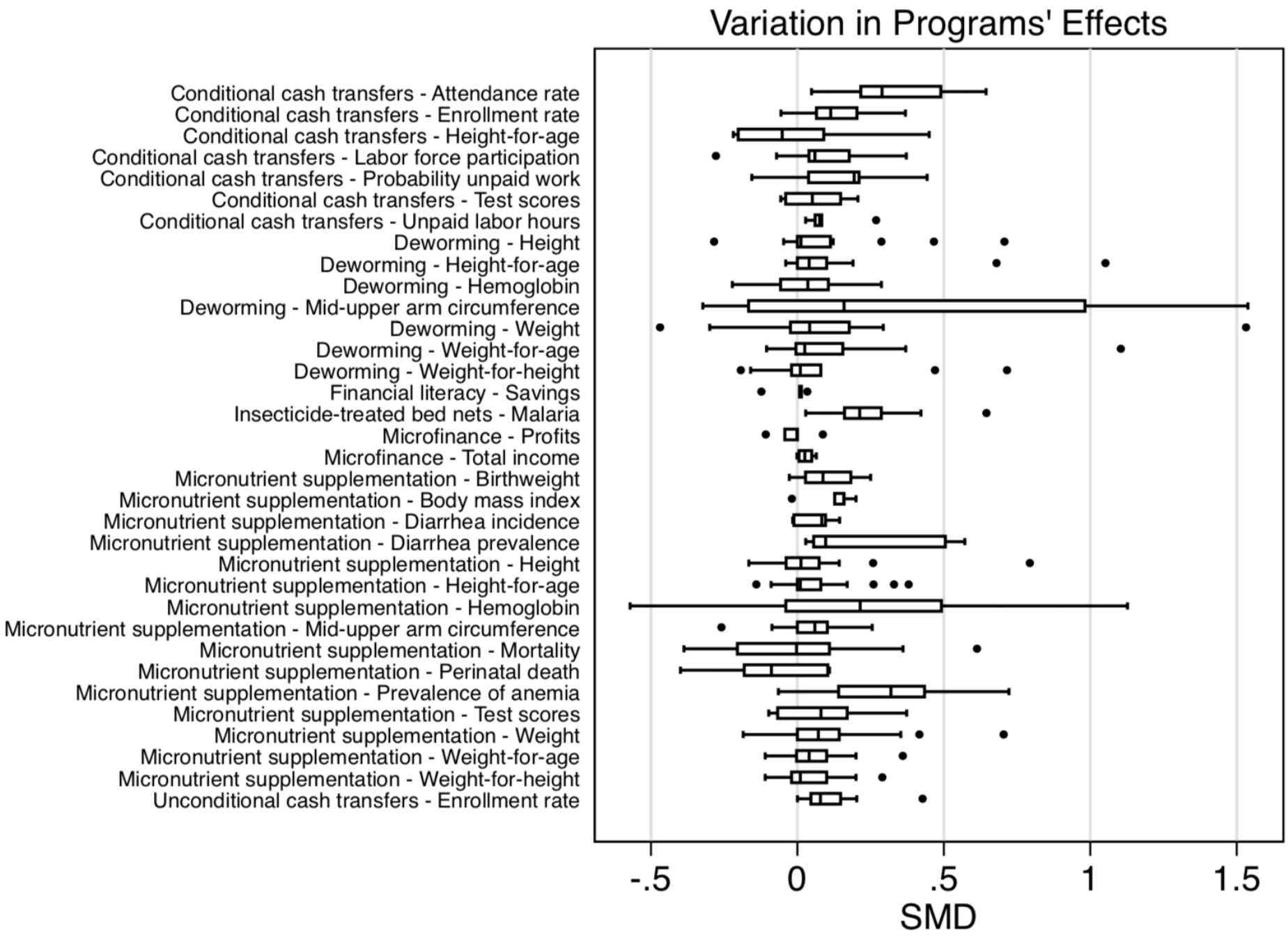
- Also need to ensure outcomes representing improvements all have the same sign (*e.g.* a decrease in disease incidence is a good thing)

Data

- When looking at ability to generalize within a set, the set is critical
- "Strict", "loose", and "broad" outcome definitions
- For generalizability, requires common outcomes
- Separate paper on specification searching and significance inflation (Vivalt 2019):
  - Not much evidence of it in the sample, especially for RCTs
  - Supports work by Brodeur et al. (2016, 2020) also finding little evidence of bias among RCTs in economics
  - Will use the subset of RCTs as a robustness check

Distribution of Outcomes Considered

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


Variation in Programs' Effects
Conditional cash transfers - Attendance rate
Conditional cash transfers - Enrollment rate
Conditional cash transfers - Height-for-age
Conditional cash transfers - Labor force participation
Conditional cash transfers - Probability unpaid work
Conditional cash transfers - Test scores
Conditional cash transfers - Unpaid labor hours
Deworming - Height
Deworming - Height-for-age
Deworming - Hemoglobin
Deworming - Mid-upper arm circumference
Deworming - Weight
Deworming - Weight-for-age
Deworming - Weight-for-height
Financial literacy - Savings
Insecticide-treated bed nets - Malaria
Microfinance - Profits
Microfinance - Total income
Micronutrient supplementation - Birthweight
Micronutrient supplementation - Body mass index
Micronutrient supplementation - Diarrhea incidence
Micronutrient supplementation - Diarrhea prevalence
Micronutrient supplementation - Height
Micronutrient supplementation - Height-for-age
Micronutrient supplementation - Hemoglobin
Micronutrient supplementation - Mid-upper arm circumference
Micronutrient supplementation - Mortality
Micronutrient supplementation - Perinatal death
Micronutrient supplementation - Prevalence of anemia
Micronutrient supplementation - Test scores
Micronutrient supplementation - Weight
Micronutrient supplementation - Weight-for-age
Micronutrient supplementation - Weight-for-height
Unconditional cash transfers - Enrollment rate
-.5
0
.5
1
1.5
SMD

Results

Table: Summary of Generalizability Measures by Heterogeneity Measures

	$\widehat{P(Sign)}$			$\widehat{\sqrt{MSE}}$			N		
	$\widehat{\tau}^2_N$			$\widehat{\tau}^2_N$			$\widehat{\tau}^2_N$		
$\lvert\widehat{\mu}_N\rvert$	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

Results

Table: Generalizability Measures by Study Quality

	$\widehat{P(Sign)}$	$\widehat{\sqrt{MSE}}$	$\widehat{\tau}^2_N$	$\widehat{I}^2_N$	$\frac{\widehat{\tau}_N}{\lvert\widehat{\mu}_N\rvert}$	$\widehat{\mu}_N$	$\widehat{s}_N$	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

Summary Statistics

- An inference about another study will have the correct sign about 61% of the time
- 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 MSE to that mean is 2.49 across intervention-outcome combinations
- Only about 6% of total variance can be attributed to sampling variance
- Modelling the variation with a mixed model can help a little, but not a lot....

Model Heterogeneity

- Generally not enough data for meta-regression
- Best-case scenario still doesn't help much

Table: Residual Heterogeneity Measures by Intervention-Outcome

Intervention-Outcome	Explanatory Variable	R^2	$\widehat{\tau^2}$	$\widehat{\tau^2_R}$	$\frac{\widehat{\tau^2}-\widehat{\tau^2_R}}{\widehat{\tau^2}}$
CCTs-Attendance rate	Baseline enrollment rate	0.43	0.0031	0.0029	0.08
CCTs-Enrollment rate	Min household non-educ. transfer	0.28	0.0010	0.0008	0.20
CCTs-Labor force particip.	Conditional on health check	0.38	0.0012	0.0013	-0.07
UCTs-Enrollment rate	Sample minimum age	0.34	0.0006	0.0006	0.04
Deworming-Height	Mebendazole dosage	0.32	0.2201	0.2097	0.05
Deworming-Height-for-age	Mix of drugs	0.32	0.0497	0.0373	0.25
Deworming-Hemoglobin	Baseline prevalence *T. Trichiura*	0.36	0.0077	0.0083	-0.07
Deworming-Weight	Baseline prevalence hookworm	0.73	0.3596	0.1153	0.68
Deworming-Weight-for-age	Baseline prevalence *T. Trichiura*	0.39	0.0114	0.0101	0.11
Deworming-Weight-for-height	Baseline prevalence hookworm	0.92	0.0191	0.0052	0.73

Potential Explanatory Factors

Table: Regression of Effect Size on Study Characteristics

	(1)	(2)	(3)	(4)	(5)
Number of	-0.013**			-0.013**	-0.011**
observations (100,000s)	(0.01)			(0.01)	(0.00)
Government-implemented		-0.081***			-0.073***
		(0.02)			(0.03)
Academic/NGO-implemented		-0.018			-0.020
		(0.01)			(0.01)
RCT			0.021		
			(0.02)		
East Asia				0.002	
				(0.03)	
Latin America				-0.003	
				(0.03)	
Middle East/North				0.193**	
Africa				(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

Potential Explanatory Factors

Table: Regression of $\widehat{\tau}^2_N$ and $\widehat{I}^2_N$ on Intervention Characteristics

	$\widehat{\tau}^2_N$			$\widehat{I}^2_N$		
	(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

CCTs on Enrollment Rates

Table: Regression of Treatment Effects on Study Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Baseline Enrollment Rates	-0.205***	-0.102***				-0.090***
	(0.05)	(0.03)				(0.03)
Enrolled at Baseline		0.001				
		(0.02)				
Not Enrolled at		0.195***				0.199***
Baseline		(0.03)				(0.02)
Number of			-0.008			
Observations (100,000s)			(0.00)			
Rural				0.038**		0.013
				(0.02)		(0.01)
Urban					-0.049***	-0.018
					(0.01)	(0.01)
Observations	249	249	145	270	270	249
R^2	0.32	0.44	0.00	0.05	0.03	0.45

Conclusions

- Impact evaluations have significant predictive power
- There remains a lot of dispersion of results
- Generalizability is modestly improved by considering explanatory variables
- Large and government-implemented projects fare worse than small, NGO/academic-implemented projects
- Interventions that have more direct causal chains fare a little better

Posterior for $\mu|\tau$

Prior: $\mu|\tau$ is uniformly distributed.

Likelihood (data): Y_i are estimates of μ with variance $\sigma_i^2 + \tau^2$.

$\Longrightarrow \mu|\tau, Y \sim N(\hat{\mu}, V_\mu)$ where

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

Posterior for τ

Prior: τ is uniformly distributed.

Likelihood (data): $Y_i \sim N(\mu, \sigma_i^2 + \tau^2)$

$$p(\tau|Y) = \frac{p(\mu,\tau|Y)}{p(\mu|\tau,Y)}$$

Numerator:

$$p(\mu,\tau|Y) \propto p(\mu,\tau)p(Y|\mu,\tau)$$

$$p(\mu,\tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

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

Putting it together:

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

Back