How Much Can We Generalize From Impact Evaluations?

Eva Vivalt

Australian National University

June 20, 2019

Research Questions

- How much can we generalize?
- Why?
  - Implementation/context differences?
  - Sampling error?
  - Specification searching/publication bias? (Separate paper)

Impact Evaluations Are Used to Predict

- Impact evaluations used to inform future work
- Results vary
- If we don't know why, don't know what will happen

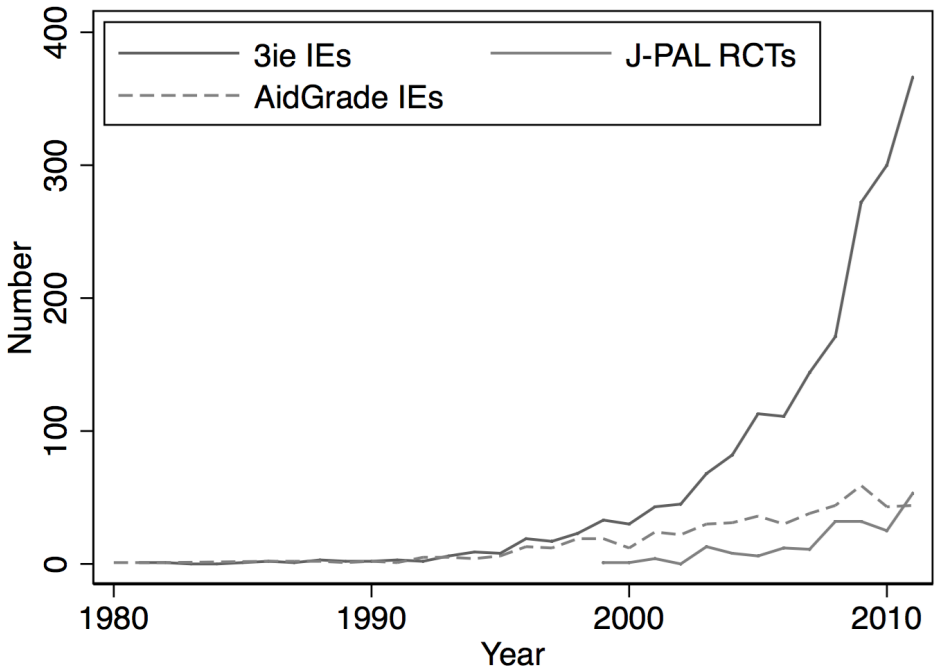
Literature

Heterogeneity in treatment effects:

- Example of same place, different effects (Bold *et al.*, 2013)
- Site selection bias (Allcott, 2012)
- 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)

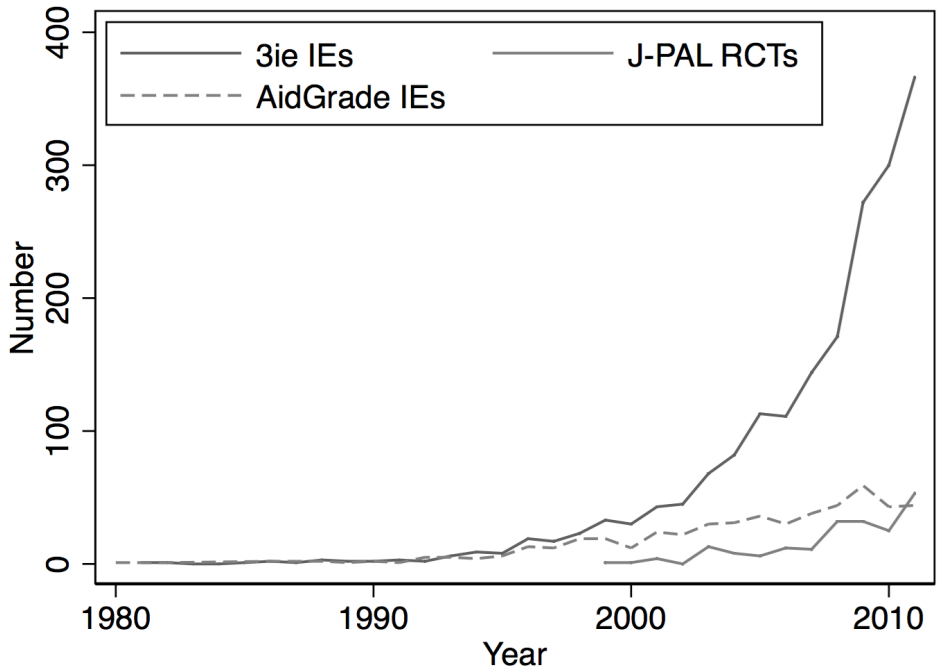
Data exists

Figure: Growth of Impact Evaluations



Data exists

Figure: Growth of Impact Evaluations



I started an organization that collects all this data.

635 IEs (474 RCTs) across 20 focused areas.

Road map

- Theory:
  - Meta-analysis models and how to estimate them
- Method:
  - Heterogeneity measures
- Data:
  - Sample selection, summary statistics
- Results:
  - Without considering study or intervention characteristics, an inference about another study will have the correct sign 67% of the time. Ratio of the $\sqrt{MSE}$ to the mean is 2.15
  - 9% of observed variation is sampling variance
  - Unlikely to be specification searching / pub bias (companion paper)
- Conclusions

Meta-Analysis Models

There are two main models used in meta-analysis: fixed effect or random effects models.

- Fixed effect:

$$Y_i = \theta + \varepsilon_i$$

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

- Random effects:

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

where θ_i is the effect size for a particular study, $\bar{\theta}$ is the mean

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:

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

where σ_i^2 is the sampling variance.

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.

Computation:

1. Simulate τ
2. Simulate μ
3. Simulate θ_i

Mixed Models

- Sometimes one wants to do moderator / mediator analysis, including explanatory variables X_n, *e.g.* linear meta-regression
- This is called a "mixed model"
- Methods are the same except involve estimating a vector of β

Measuring Generalizability

- How to define generalizability?
- How to relate it to heterogeneity measures?

Heterogeneity Measures

- Two classes of measures:
  - Variation
  - Proportion of variation that is systematic

Heterogeneity 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 that is not sampling error
  - 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)

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 (sign) and Type M (magnitude) errors
- 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

- In particular, 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 variables in the Bayesian model: τ^2, μ, σ_i^2 (or I^2)
- So we can estimate values for these parameters and then talk of inference errors of sign and magnitude
- 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

Data

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

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

Data

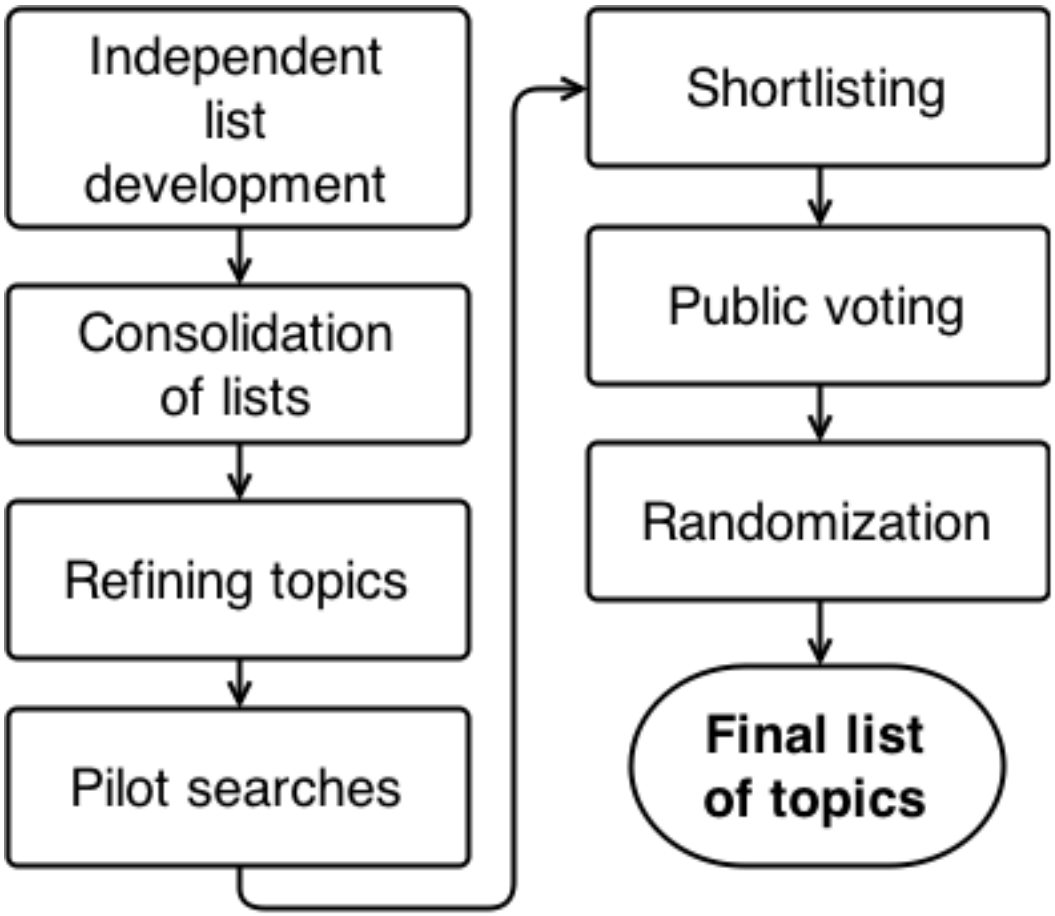
- Any impact evaluation attempting to measure counterfactual is included
- Published papers and working papers both included
- 85 fields, including 13 for identifying information (author name, publication year, program name, *etc.*) converted to 89 variables. Additional topic-specific fields
- Heterogeneity in programs and samples. Program and sample characteristics coded but frequently too sparse to use
- Double-entry coding for everything

Process

- Selection of interventions
- Search
- Screening
- Data extraction

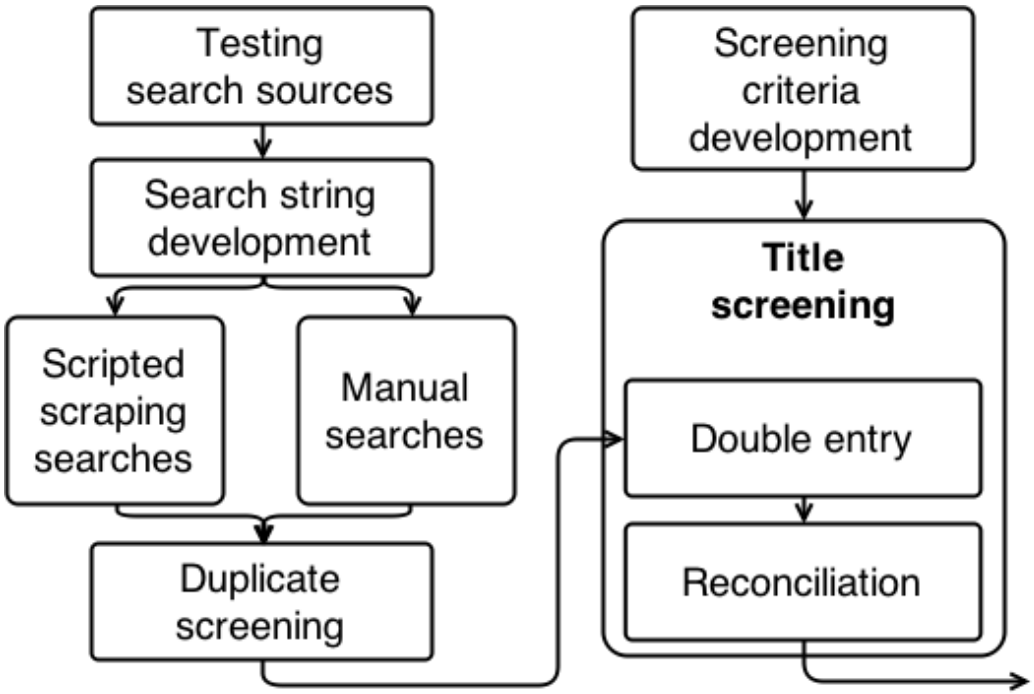
Process Diagram

Figure: Topic Selection



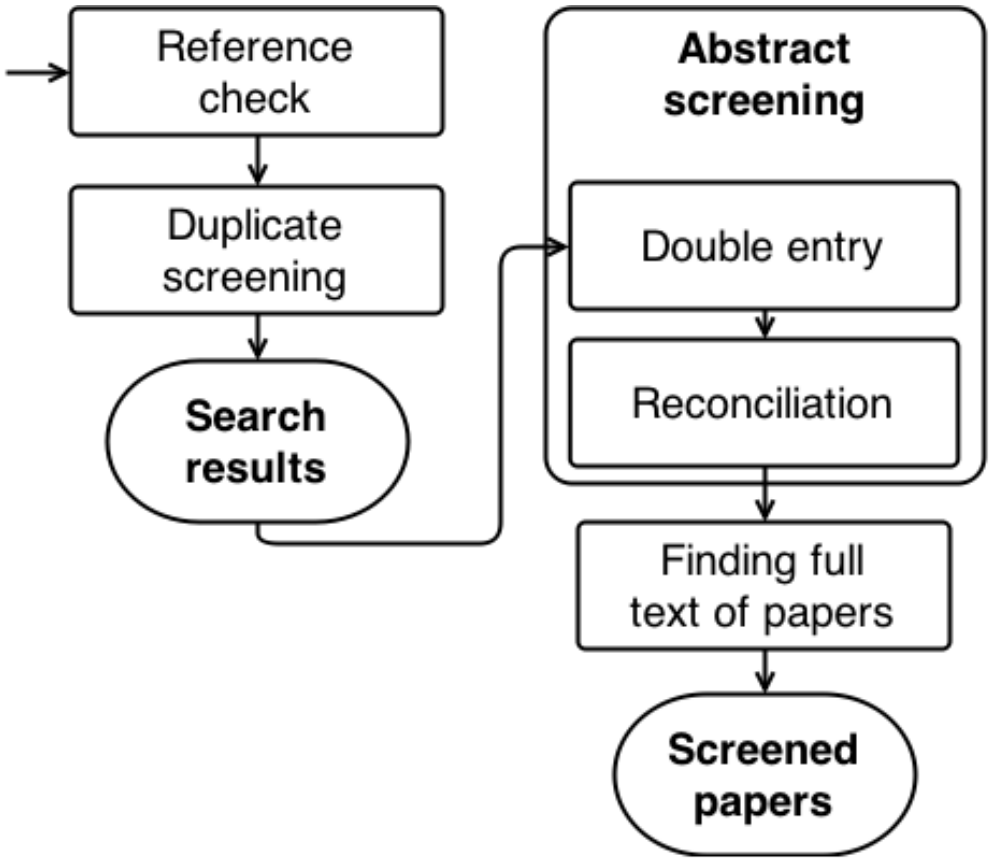
Process Diagram

Figure: Search and Screening, Part 1



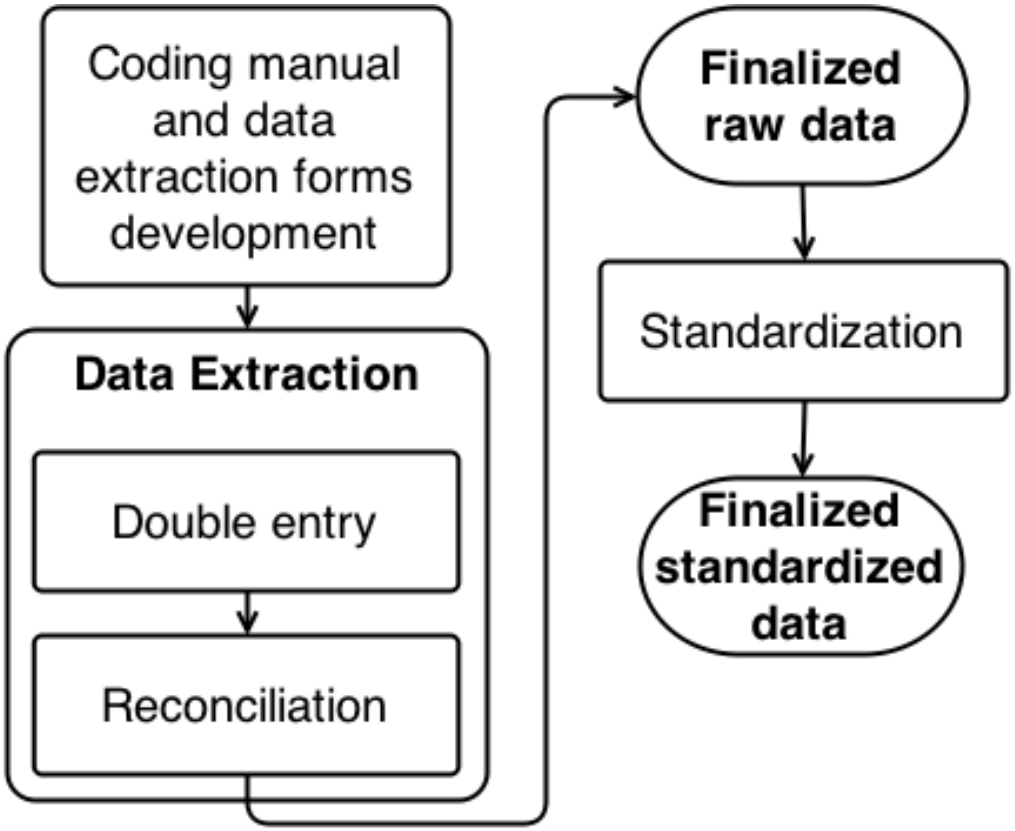
Process Diagram

Figure: Search and Screening, Part 2



Process Diagram

Figure: Data Extraction



Data

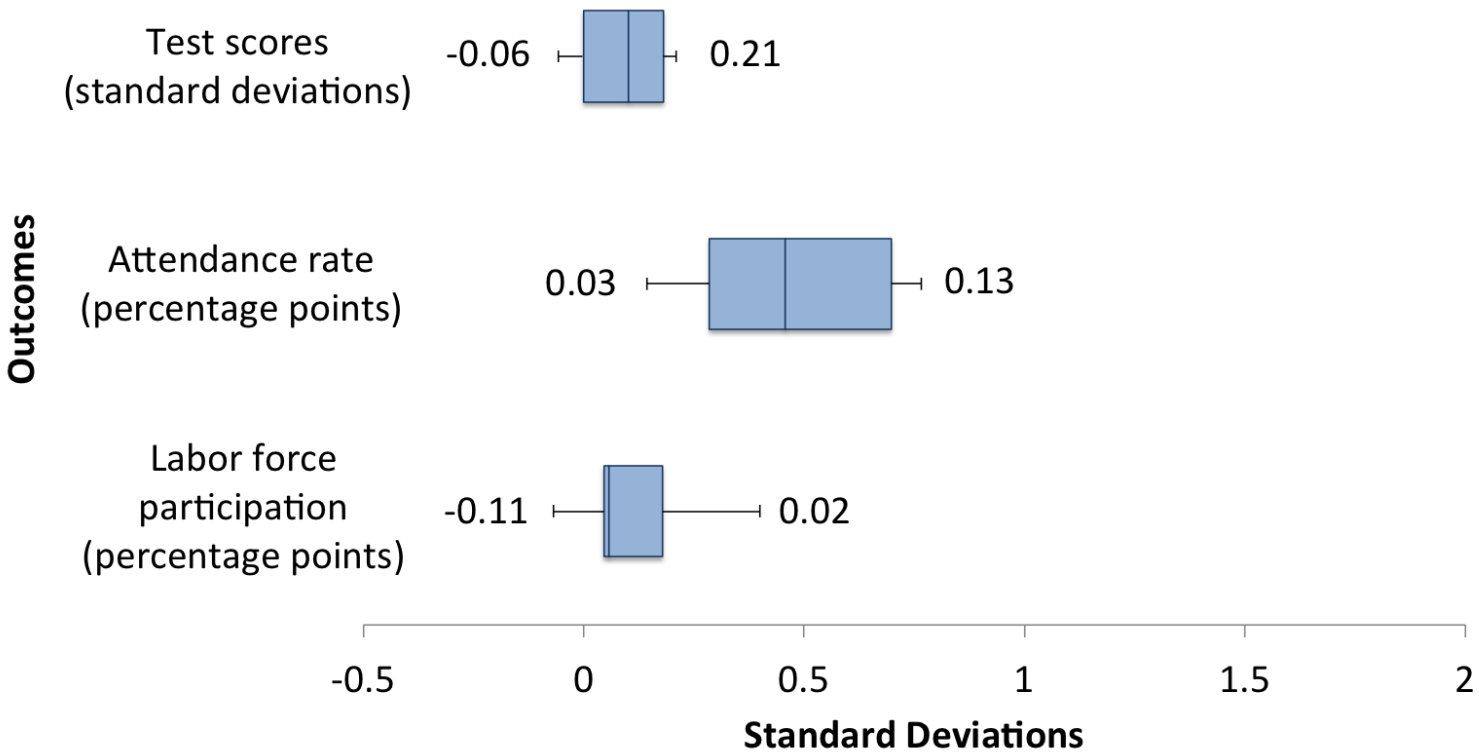
- For **a subset** of analyses, need to standardize effect sizes:

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

- Also need to ensure outcomes representing improvements all have the same sign (*e.g.* a decrease in disease incidence is a good thing)
- In general, I try to represent results in raw units and report disaggregated results

Standardized and Transformed vs. Raw Data

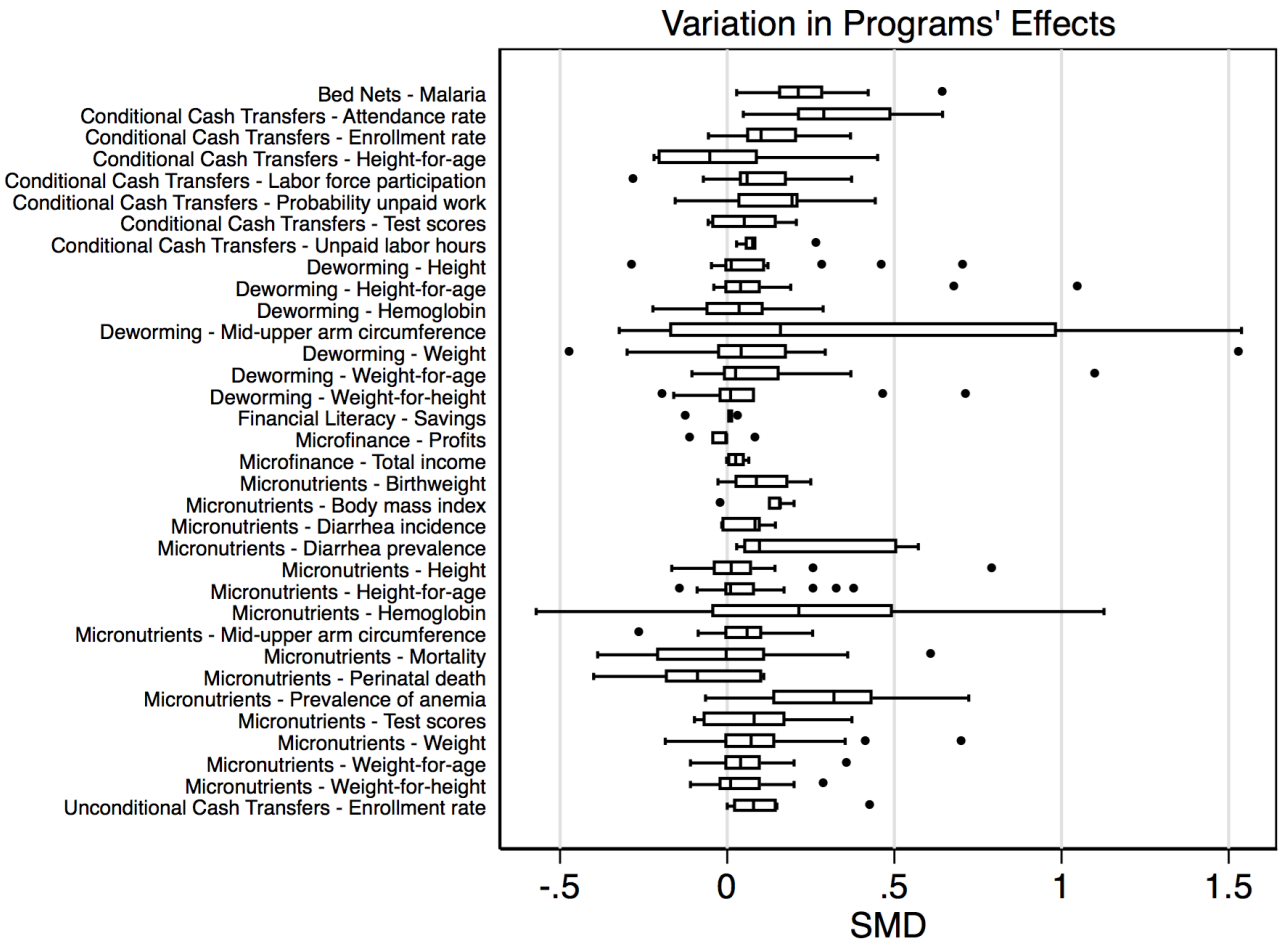
Figure: Selected Outcomes for Conditional Cash Transfers



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 and reduced to one per paper): 649 results across 277 papers if retaining intervention-outcome combinations covered by at least two papers; 576 results across 251 papers if retaining intervention-outcome combinations covered by at least three.
- For specification searching and publication bias: 11,970 results from 584 papers.

Dispersion of Estimates



Results

- Median probability that the sign of a similar study would be correctly predicted: 67%
- 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 - malaria; CCTs - enrollment rates; SMS reminders - attendance rates
- Microfinance and financial literacy training only slightly better than 50% for most outcomes
- Median $\frac{\sqrt{MSE}}{|\hat{\mu}|}$: 2.15

Results

- What kinds of intervention-outcomes did particularly well?
- Consider $\frac{\hat{\tau}_N}{|\hat{\mu}_N|}$:
  - Some of the lowest values are for conditional cash transfers and health-related interventions
  - Highest values for financial interventions, *i.e.* microfinance and financial literacy training

Summary Table

	$\widehat{P(Sign)}$			$\widehat{\sqrt{MSE}}$			N		
	$\widehat{\tau}^2_N$			$\widehat{\tau}^2_N$			$\widehat{\tau}^2_N$		
$\lvert\widehat{\mu}_N\rvert$	Low	Med	High	Low	Med	High	Low	Med	High
Low	0.692	0.539	0.526	0.08	0.26	0.46	14	4	1
Med	0.769	0.617	0.528	0.11	0.29	0.56	4	10	5
High	0.982	0.813	0.661	0.20	0.30	50.43	1	5	13

Modeling Heterogeneity

1. Across intervention-outcomes
2. Within intervention-outcomes

OLS 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

OLS of $\widehat{\tau}^2_N$ on Study Characteristics

	(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

OLS of of $\widehat{I^2_N}$ on Study Characteristics

	(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

OLS 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

Within-Intervention-Outcome

- Select the single best-fitting explanatory variable, maximizing R^2
- Use that variable in a mixed model to "explain" heterogeneity
- Calculate residual heterogeneity measures

Residual Heterogeneity Measures by Intervention-Outcome

Intervention	Outcome	R^2	τ^2	τ^2_R	I^2	I^2_R	N
CCTs	Attendance rate	0.43	0.0031	0.0029	0.878	0.857	8
CCTs	Enrollment rate	0.28	0.0010	0.0008	0.961	0.952	36
CCTs	Labor force participation	0.38	0.0012	0.0013	0.939	0.944	10
UCTs	Enrollment rate	0.34	0.0006	0.0006	0.844	0.848	10
Deworming	Height	0.32	0.2201	0.2111	0.942	0.940	13
Deworming	Height-for-age	0.32	0.0500	0.0373	0.989	0.986	13
Deworming	Hemoglobin	0.36	0.0078	0.0082	0.645	0.657	11
Deworming	Weight	0.73	0.3587	0.1153	0.995	0.984	9
Deworming	Weight-for-age	0.39	0.0114	0.0101	0.966	0.960	8
Deworming	Weight-for-height	0.92	0.0189	0.0053	0.910	0.604	5

Conclusions

- Impact evaluations are informative about one another, yet there remains a lot of dispersion of results
- Government-implemented projects fare worse than NGO/academic-implemented projects
- Larger projects obtain smaller effect sizes
- Incentivized and health interventions may have more generalizable results
- Sampling variance does not contribute much to overall heterogeneity
- Generalizability is modestly improved by considering explanatory variables