Experimental Methods: Lecture 7

Generalizaton and Meta Studies

Raymond Duch

June 6, 2023

DPIR University of Oxford

Road Map to Lecture 7

- External Validity: Overview
- External Validity: Dimensions
- External Validity: Why does it matter?
- Meta Studies and External Validity

External Validity: Overview

Internal Validity Revisited

Internal validity

- Answer the question: do the experimental data allow for correct causal inferences?
- Depends fundamentally on experimental design and data analysis
  - Random assignment
  - Excludability
  - Non-interference
  - Measurement error
- Traditionally the main concern of experimental research in political science

The internal validity of experiments is what we have mainly focused on so far.

External Validity

- Researchers and policy makers are often interested in drawing generalised knowledge from experimental results
- Does the impact of women's leadership on policy decisions, as measured in Village Councils in West Bengal, hold in all rural India?

External Validity

External validity

- Answer the question: to what extent internally valid experimental results apply beyond their immediate objects of investigation?

Generalizability vs Transportability

Inferences of external validity can take two forms

Generalizability

- Extent to which inferences drawn from a given study's sample S apply to the broader population P
  - where $S \subseteq P$

- Can you think of an example?

Generalizability vs Transportability

Transportability

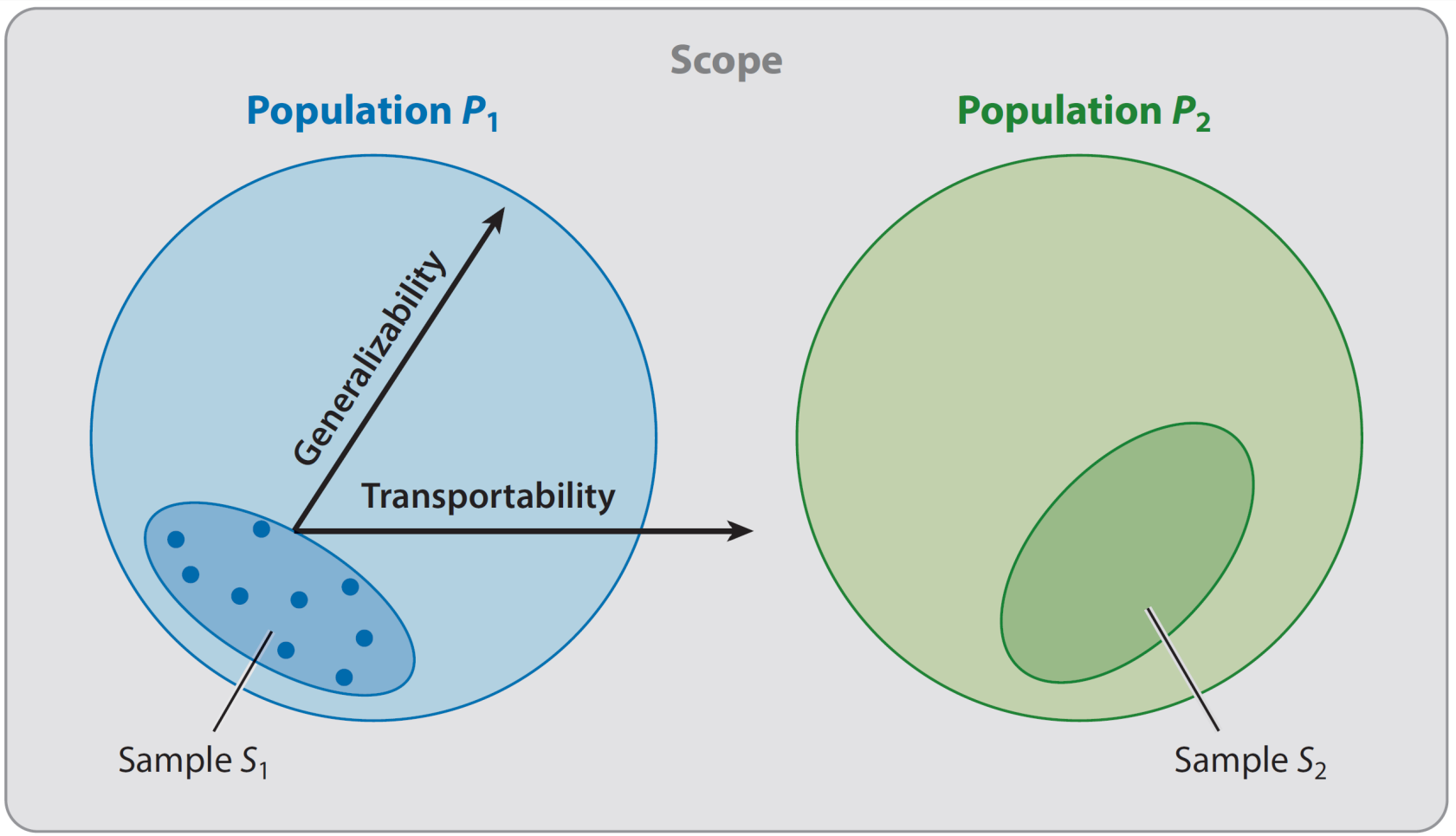
- Extent to which inferences drawn from a given study's sample S apply to another target population P
  - where $S \nsubseteq P$

- Do behavioural deviations from the *Homo economicus* model observed in WEIRD populations apply in other cultures?

Scope, Populations, and Samples

- The *scope* of a study defines the study's applicability and limitations
- It encompasses the definitions of key parameters
  - theoretical population(s)
  - accessible population(s)
  - associated samples

Scope, Populations, and Samples



External Validity: Dimensions

Cronbach Shapiro (1982): UTOS framework

- Inferences of external validity involve the analysis of several dimensions
- UTOS framework focuses on
  - Units
  - Treatments
  - Outcomes
  - Settings

Units

- Unit-specific external validity: inferences estimate for which population units sample inferences hold

Treatments

- External validity of a treatment
  - Do inferences hold across different operationalizations of the main explanatory variable

- Treatment variable must have construct validity
  - Treatment must be operationalized to correspond to theoretical concept of interest
  - Construct often sacrificed to convenience

Outcomes

- The same theoretical construct (e.g. social trust) of interest can be operationalize using different measures
- Do inferences hold across different operationalizations?
- Outcome-related external validity requires some form of mundane realism

Settings

- Environments in which a study's data are generated (e.g. laboratory, a country, a school)
- Different settings potentially yield different levels of external validity
- Particularly true when comparing laboratory, survey, and field experiments

Time

- Treatment effects, including those for relevant subgroups, change
- Composition of the population of interest is not static

- All relevant confounders are not identifiable and measurable

Mechanism

- mechanisms can be considered to be mediators occurring after a treatment and before an outcome
- mechanisms take on many other forms including constraints, equifinality, and interactions

Findley et al. (2021): M–STOUT framework

Table 1 Examples of M–STOUT

Dimension[a]	Study A[b]	Study B[b]	Target inference[c]
Mechanisms	Women's empowerment	NA	Youth empowerment
Settings	Survey experiment in Liberia	TSCS regression of African countries	Field experiment in Guinea
Treatments	CDD projects	Any aid projects	Direct budget support
Outcomes	Self-reported social trust	WVS indicators of social trust	Results of a social trust game
Units	Individuals within villages	African countries	Individuals within counties
Time	Year of 2000	Years of 2000–2020	Year of 2020

[a]Each M–STOUT element is listed in this column.
[b]The corresponding elements in a study about Liberia (Study A) and a TSCS study about African countries, 2000–2020 (Study B).
[c]Examples of possible external validity targets.
Abbreviations: CDD, community-driven development; NA, not applicable; TSCS, time series cross-sectional; WVS, World Value Survey.

External Validity: Why does it matter?

External Validity: Why does it matter?

- Consider an hypothetical study on the effect of an aid program on social trust
- We can determine the in-sample average treatment effect (*SATE*) using the difference-in-means estimator

$$\hat{\delta_S} = \hat{y}(1) - \hat{y}(0)$$

- We can re-write it in terms of the effect in the population of interest (*PATE* for generalizability and *TATE* for transferability) plus potential biases of the estimator

$$\hat{\delta_S} = \delta_P + b_S1 + b_S2 + b_P + b_V$$

Sources of Bias

What sources of bias should we consider?

$$\hat{\delta_S} = \delta_P + b_S1 + b_S2 + b_P + b_V$$

Internal validity biases

- Selection bias into treatment (b_S1)
- In-sample treatment effect heterogeneity (b_S2)

External validity biases

- Sample selection bias (b_P) - difference in treatment effects between sample and non-sample units weighted by proportion of excluded units in P
- Variable selection bias (b_V) - PATE/TATE with variables of interest minus the PATE/TATE with variables at hand

Sources of Bias

- Proper randomisation of units to treatment and control groups usually ensures that $b_S1 = b_S2 = 0$
- However, an unbiased estimate of the *SATE* is potentially highly misleading if *SATE* is very different from *PATE*/*TATE* of interest
- Which strategies can we implement to formulate robust/credible external validity inferences?

Findley et al. (2021): Model Utility

- Theoretical model guiding the application of effects, findings, and inferences derived from study samples to the target population
- Focus on the mechanisms explaining study findings, rather than point estimates
- Specifically articulates the causal principles operating in the sample and in the target population
    - Critically, causal principles characterize the interactions of a cause with the underlying context
- Clarifies the level of abstractions of the mechanism
    - Treatments and variables of interest often vary across contexts operationalizing similar theoretical constructs

Meta Studies and External Validity

Meta Studies

- Can meta studies help to draw external validity inferences?
- Aggregate causal evidence from studies varying in samples, contexts, or times

Slough & Tyson (2023)

TABLE 1 Meta-Analyses in Political Science and General Science Journals

	Stated Meta-Analysis Motivation			Component Study Design		Estimator		
	Generalizability	Precision	Literature Synthesis	Experimental	Observational	RE	FE	Other
Prospective Meta-Analysis of Harmonized, Original Studies								
Dunning et al. (2019)	✓			✓			✓	✓
de la O et al. (2021)	✓			✓		✓		
Slough et al. (2021)	✓	✓		✓		✓		
Blair et al. (2021)	✓			✓		✓		
Coppock, Hill, and Vavreck (2020)	(✓)	✓		✓		✓		
Retrospective Meta-Analysis based on Secondary Analysis of Existing Studies								
Blair, Christensen, and Rudkin (2021)			✓		✓	✓	✓	
Blair, Coppock, and Moor (2020)			✓		✓†	✓		
Eshima and Smith (2022)			✓	✓		✓		
Incerti (2020)			✓	✓		✓	✓	
Godefroidt (2021)	✓		✓		✓*	✓		
Kertzer (2020)	✓		✓		✓‡	✓		
Schwarz and Coppock (2022)	✓		✓	✓		✓		
Meta-Analysis based on Secondary Analysis and Original Studies								
Kalla and Broockman (2018)	(✓)		✓	✓		✓		

Vivalt (2020)

- Impact evaluation results are widely cited in reports generated for policymaking
- Often shared without much information about context, study design or even standard errors
- How much do results truly vary?
- Are there characteristics of studies that predict generalizability?

Vivalt (2020)

- Vivalt (2020), How Much Can We Generalize From Impact Evaluations?
- Measure inter-study effect heterogeneity τ of set of studies S
- Given S, use τ to estimate
  - likelihood of predicting the *sign* of the true effect of a similar study in another context
  - the *magnitude* of the error associated with that prediction

Vivalt (2020) - Method

Two general modelling approach in meta studies

- Fixed-effect models - effect heterogeneity across implementations of same intervention due only to sampling error

  $$Y_i = \theta + \epsilon_i$$

  where Y_i point estimate in study i, θ the true effect, and ϵ_i the error term
- Random-effect models

  $$Y_i = \theta_i + \epsilon_i$$

  where the true effect θ_i is allowed to vary across implementations

Inverse-Variance Method

- Common approach for estimating random-effect models in meta studies
- Aims to minimise the variance of the weighted mean
- Weights each study's point estimate by the inverse of its variance

$$\hat{y} = \frac{\sum_{i=1}^{N} y_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2}$$

Bayesian Hierarchical Random-Effects Model

A Bayesian approach defines

$$\theta_i \sim N(\mu, \tau^2)$$

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

where

- Y_i is the study's point estimate
- μ is the grand mean
- τ^2 is the true inter-study variance
- σ_i^2 is the sampling variance of the error, assumed to be normally distributed for large samples

Mixed Model

Explanatory variables can be added to the model, which takes the form

$$Y_i = \alpha + X_i\beta + \epsilon_i + u_i$$

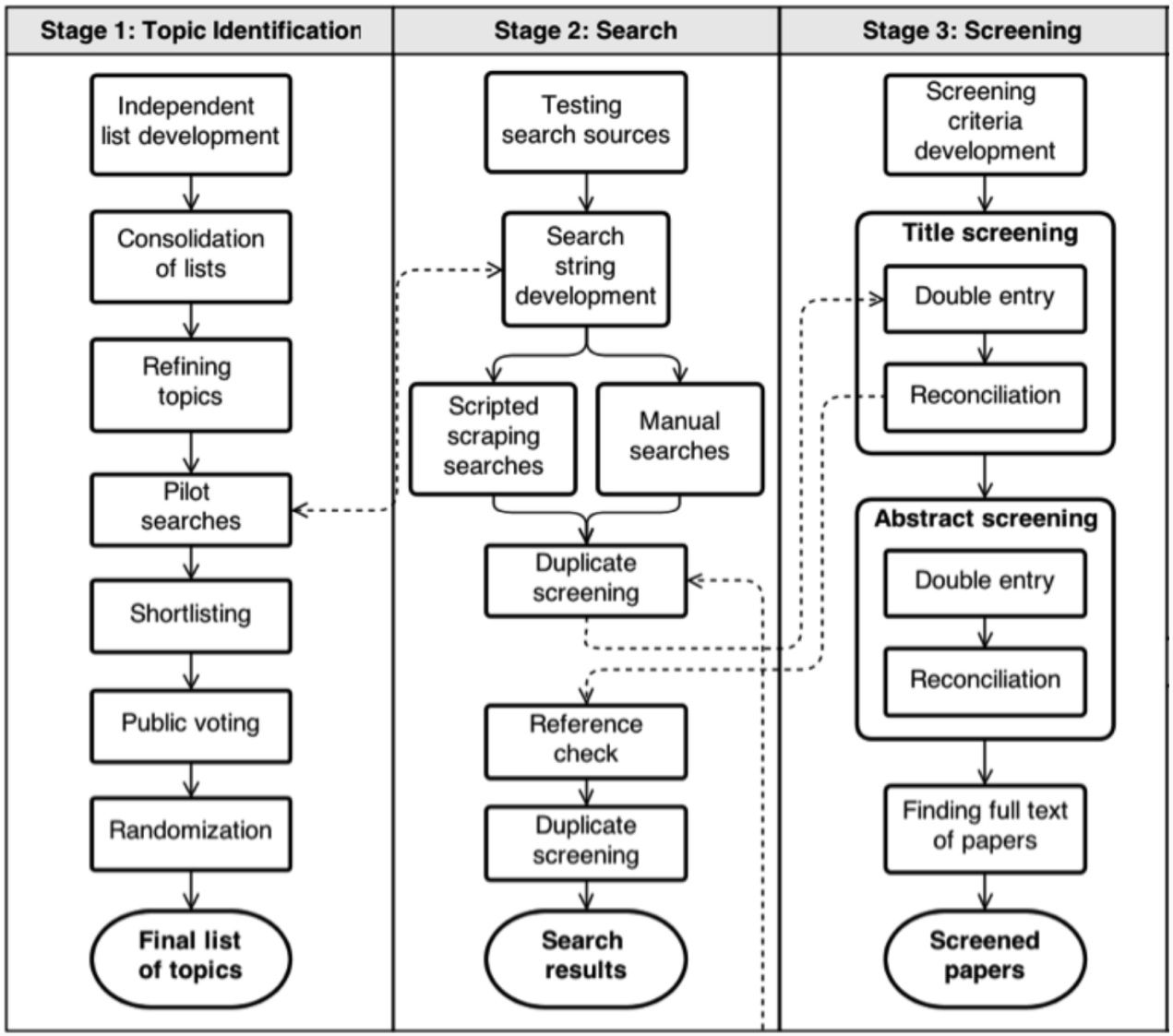
where

- $\epsilon_i \sim N(0, \tau^2)$ captures the true unexplained variance between studies
- $u_i \sim N(0, \sigma_i^2)$ captures the sampling error

Vivalt (2020): Data

- Database of impact evaluation results collected by AidGrade, non-profit focusing on meta-analysis of impact evaluations
- Data set of 15,024 estimates from 635 papers on 20 types of interventions in international development

Vivalt (2020): AidGrade Data Collection Process

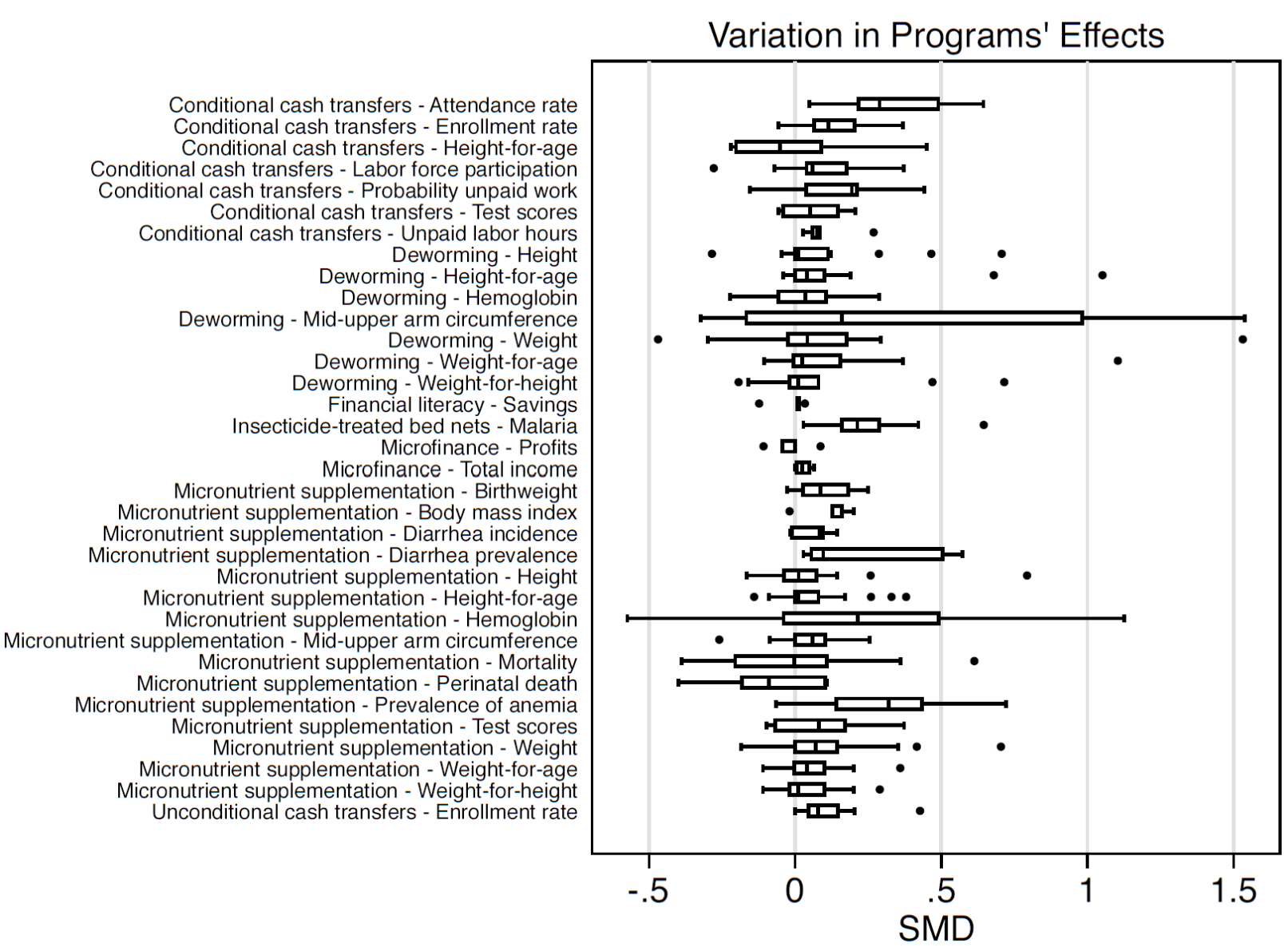


Vivalt (2020): Results

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

Vivalt (2020): Results



Vivalt (2020): Results

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

Vivalt (2020): Results

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. $\hat{\tau}^2_N$, $\hat{I}^2_N$, $\hat{\tau}_N/|\hat{\mu}_N|$ and $\hat{\mu}_N$ likewise present the average estimate for each parameter. $\hat{s}_N$ estimates a common sampling error for each intervention-outcome using Higgins and Thompson's approximation. It is important in estimating $\hat{I}^2_N$ 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 $\hat{I}^2_N$ 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.

Vivalt (2020): Results

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.

Vivalt (2020): Results

Some descriptive statistics

- An inference about another study will have the correct sign about 61% of the time

- In prediction of treatment effect of a similar study based on mean treatment effect in an intervention-outcome combination, the median ratio of the MSE to that mean is 2.49 across intervention-outcome combinations

Vivalt (2020): Results

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.

SCIENCE ADVANCES | RESEARCH ARTICLE

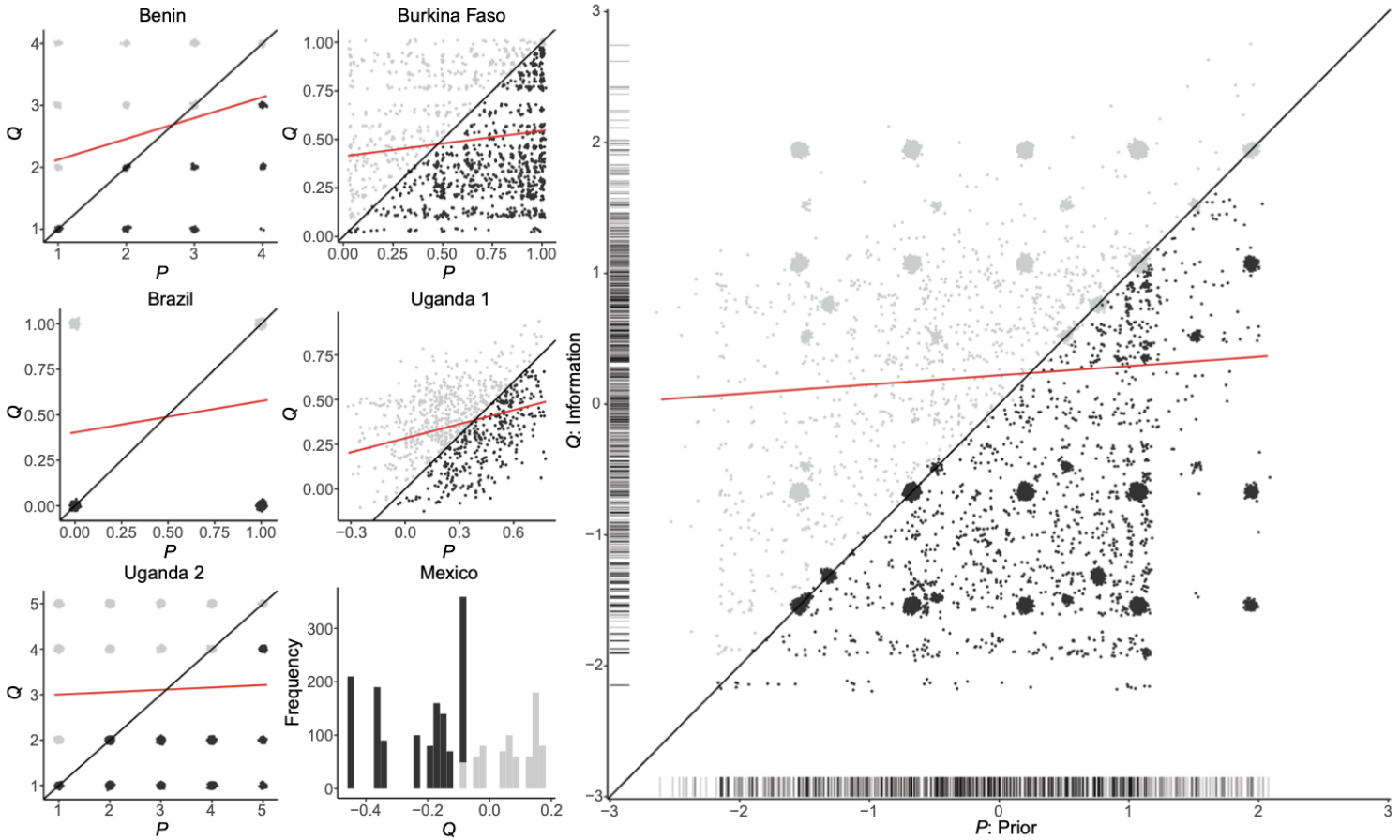


Fig. 1. Prior beliefs and politician performance. The figure plots performance information (*Q*) against prior beliefs (*P*) in each of the studies (**left**) and across all studies (**right**). Voters are in the good news group (gray) if information exceeds priors ($Q > P$) or if it confirms positive priors ($P = Q$, and Q is greater than median); otherwise, they are in the bad news group (black). On the right side, *P* and *Q* are standardized with a mean of 0 and an SD of 1 in each study. The density of the dotted areas is proportionate to the number of voters at each value of *P* and *Q*; for the pooled analysis, the rugs along the horizontal and vertical axes indicate the distribution of values. The Mexico study lacked a preintervention survey; thus, we determine the good news and bad news groups according to whether *Q* is greater than the median. The red lines indicate the linear fit between priors and information. For the pooled analysis, the slope of the fit is 0.071; the correlation is 0.053.

Metaketa

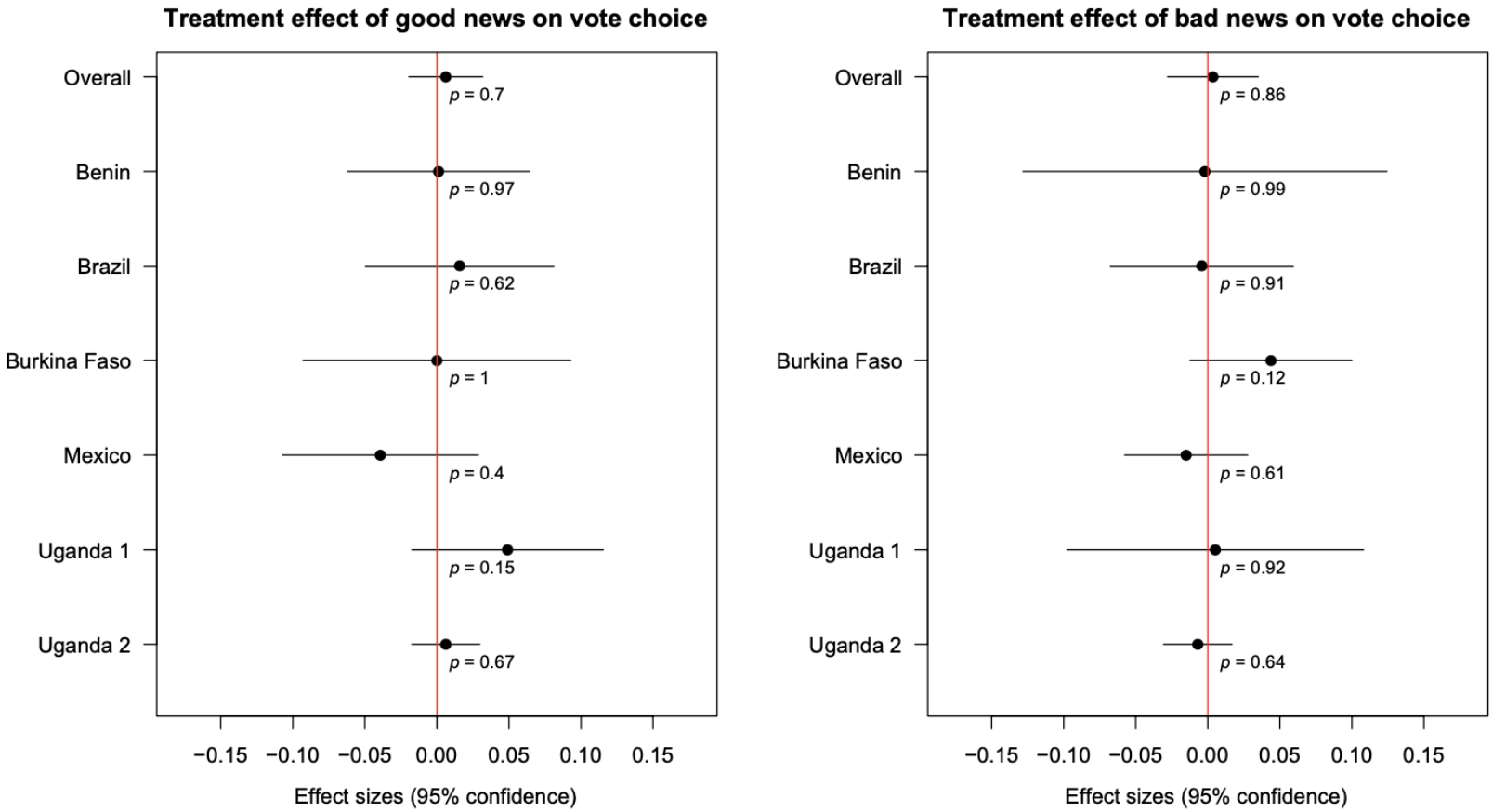


Fig. 2. Meta-analysis: Country-specific effects on vote choice. Estimated change in the proportion of voters who support an incumbent after receiving good news (**left**) or bad news (**right**) about the politician, compared to receiving no information. Unadjusted estimates. For estimating the average of the study-specific effects (top row), each study is weighted by the inverse of its size. Horizontal lines show 95% CIs for the estimated change. Entries under each estimate show *p*-values calculated by randomization inference. In all cases, the differences are close to zero and statistically insignificant.