Implementation Matters: Generalizing Treatment Effects in Education

Rachael Meager (UNSW)
joint work with Noam Angrist (Oxford, Youth Impact)

AAC 2023

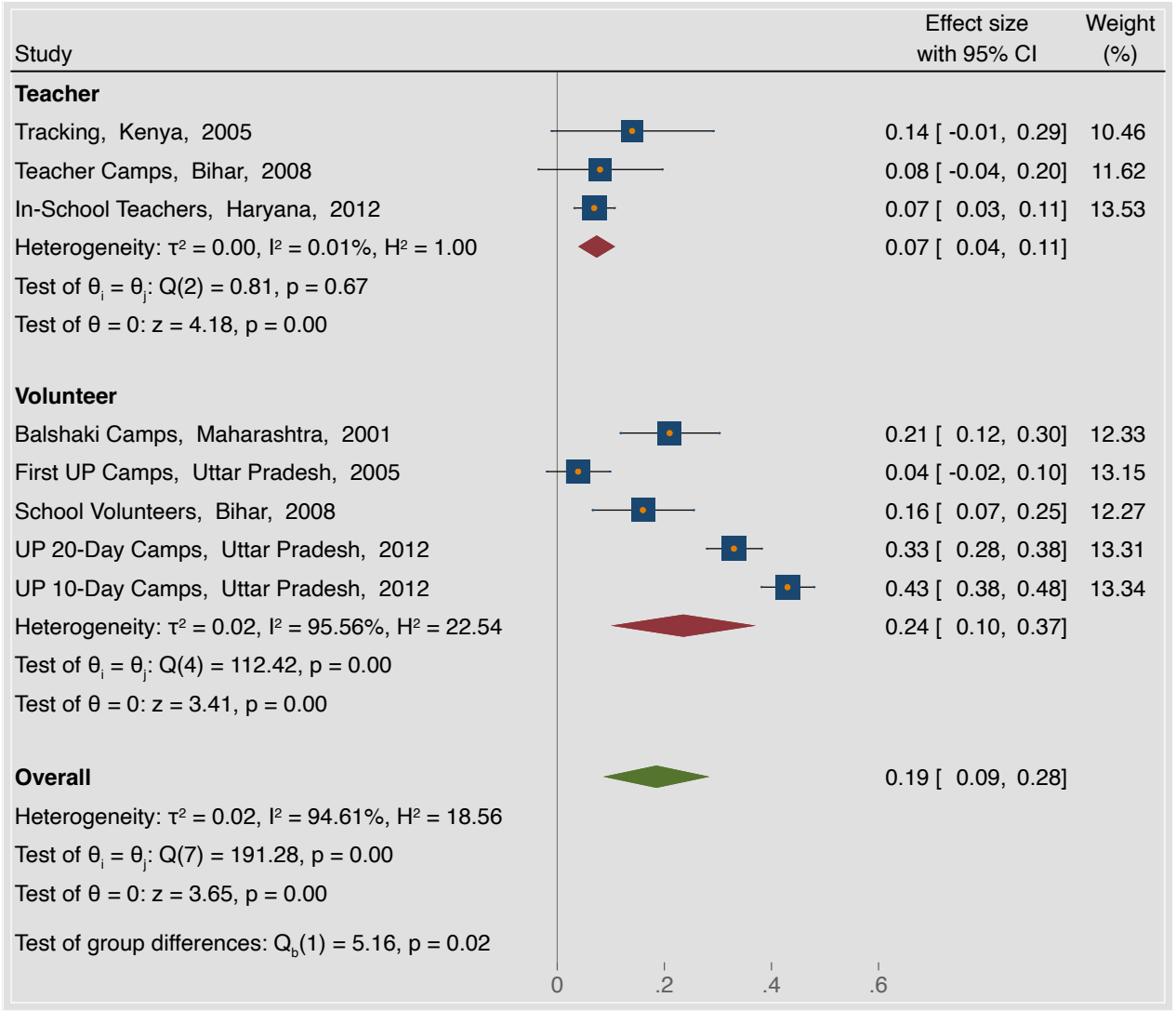
This paper studies targeted instruction

- We study targeted instruction; "teaching at the right level" (TaRL).
- Popularised and rigorously evaluated in various developing contexts by Banerjee, Duflo and others – but treatment effects vary by an order of magnitude.
- We aggregate evidence from these studies using frequentist and Bayesian methods to study this heterogeneity.
- We find that accounting for differences in implementation is essential to generalise treatment effects across settings.
- We run a new RCT showing the relationship is causal – implementation can be improved and this increases impact.

Data & Context

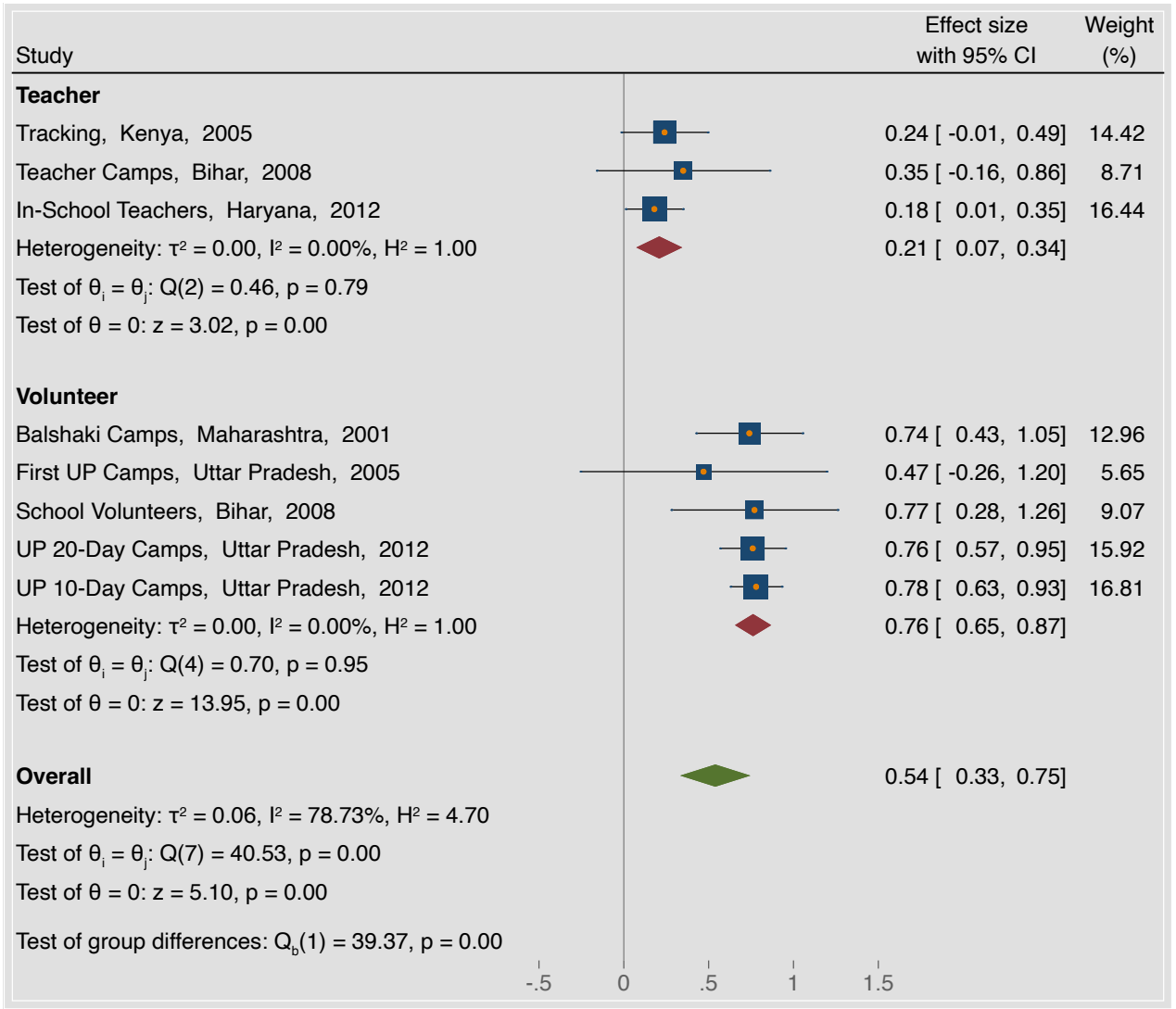
- 8 existing TaRL evaluations in 3 papers across 5 contexts: Banerjee et al. (2007) in Maharashtra, Duflo et al (2011) in Kenya, Banerjee et al. (2017) in Bihar, Uttar Pradesh, and Haryana.
- Programs differ in baseline student attainment, takeup rate among students (from 8% to 90%), taught by teachers or volunteers.
- First we consider intention-to-treat (ITT) and treatment-on-treated (TOT) separately.
- Second we run Bayesian meta-regression of effects on variety of factors.
- Main outcome: average literacy and numeracy from a test like ASER.

ITTs: big effects but heterogeneous for volunteers



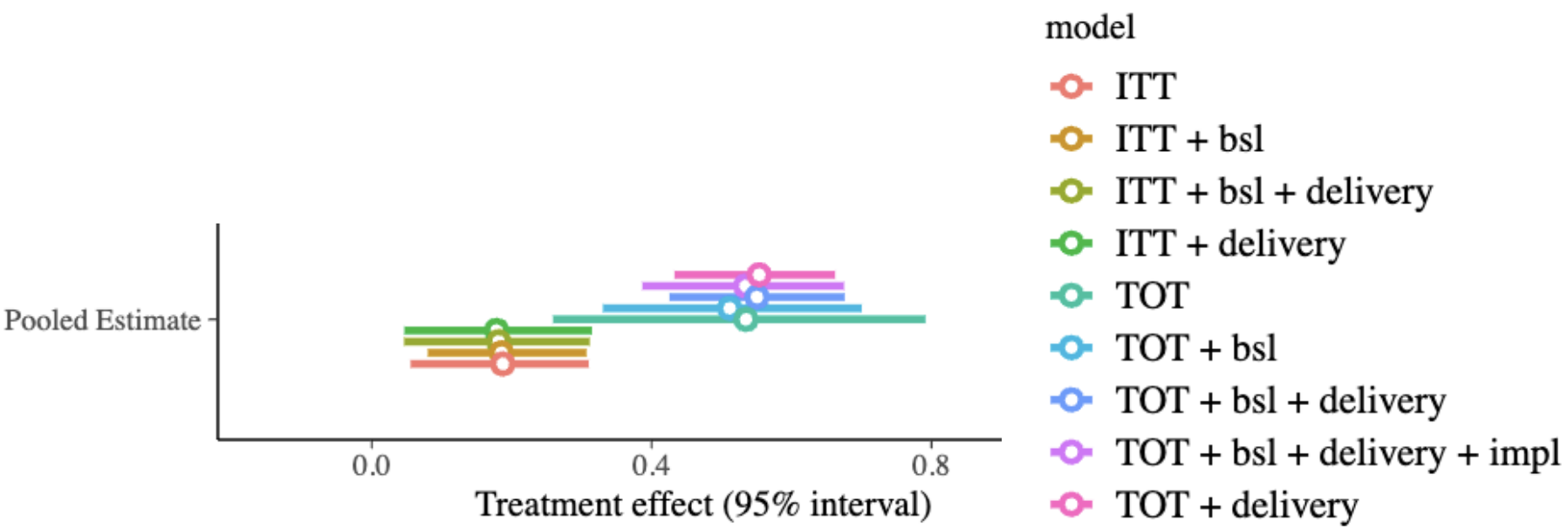
Random-effects REML model

TOTs: volunteer effects now larger + generalise



Random-effects REML model

Bayesian Meta-Regression confirms delivery matters



All regression coefficients are heavily regularised to zero. Precision of TOT result using only delivery variable is indicative.

But these results are difficult to interpret

- We found: baseline learning and country context appear less important than program takeup and delivery.
- Suggests implementation factors drive differences in results.
- Problem 1: The analysis does not account for uncertainty in takeup measures or program fidelity.
  $\rightarrow$ Solution: new joint model of uncertainty on all variables
- Problem 2: These are just correlations
  $\rightarrow$ Solution: new RCT in Botswana to vary implementation.

New joint model of effects & implementation

- Implementation is a (multi-dimensional) random variable that can be defined and measured.
- It also has a specific relationship to impact: it dials the treatment effect up or down – ie. multiplicative.
- We can measure implementation variables, but we measure them with noise. Failing to account for this could bias results.
- Jointly aggregating the information on the effects and the implementation is possible with a bit of new theory work.

Latent Effects vs Realised Effects

- Formally, we consider a set of program contexts indexed by $j = 1, 2, 3...J$ and define three relevant objects.
- The implementation factor (m-factor), denoted $m_j \in [0, 1]$ for a setting j, is the extent or proportion to which the program was effectively implemented in setting j.
- The Latent Treatment Effect LTE_j, denoted $\theta_j \in \mathbb{R}$ for a setting j, is the impact achievable when the program is fully implemented.
- The Realized Treatment Effect RTE_j is the observed impact of the program in setting j, defined as:

$$RTE_j \equiv m_j \theta_j$$

Latent Effects can be identified if m is measured

- Latent Treatment Effects matter just as much if not more than Realised Treatment Effects – especially for generalizability to future interventions.
- If implementation m is not measured (with error is fine), the Latent TEs are not identified from the Realised TEs and nor is their variation.
- We now build a Bayesian hierarchical model that decomposes each RTE_j into m_j and LTE_j while aggregating.
- We consider takeup and program fidelity as m_{1j}, m_{2j}.

Average Latent ITT of TaRL approach TOT

Table: Model with Takeup as Implementation factor: Posterior Distribution on Effects

	mean	2.5%	25%	50%	75%	97.5%	Rhat
Panel A: Latent Treatment Effects (All)							
Hypermean	0.418	0.231	0.361	0.414	0.473	0.616	1.002
HyperSD	0.207	0.068	0.136	0.188	0.256	0.464	1.002
Panel B: Latent Treatment Effects (Teacher)							
Hypermean	0.239	-0.104	0.154	0.223	0.305	0.697	1.031
HyperSD	0.235	0.005	0.060	0.142	0.312	0.922	1.021
Panel C: Latent Treatment Effects (Volunteer)							
Hypermean	0.486	0.166	0.420	0.474	0.554	0.809	1.012
HyperSD	0.233	0.017	0.087	0.164	0.296	0.930	1.006

Note: This inference is generated by $J = 7$ studies. Rhat is a diagnostic criterion for MCMC convergence with multiple chains in which a value close to 1 indicates good mixing. We use the posterior median as our preferred point estimate per the simulations in our appendix.

Average Latent TOT matches largest effects in literature

Table: Model with Fidelity and Takeup as Implementation Factors: Posterior Distributions of Effects

	mean	2.5%	25%	50%	75%	97.5%	Rhat
Panel A: Fidelity on TOT							
Hypermean	0.846	0.379	0.734	0.834	0.935	1.410	1.011
HyperSD	0.392	0.008	0.091	0.220	0.512	1.673	1.013
Panel B: Fidelity and Takeup Jointly							
Hypermean	1.200	0.126	0.998	1.140	1.356	2.324	1.009
HyperSD	0.782	0.012	0.133	0.368	0.949	4.304	1.014

Note: This inference is generated by J =3. Rhat is a diagnostic criterion for MCMC convergence with multiple chains in which a value close to 1 indicates good mixing. Panel B should be treated as suggestive because model performance measured by RMSE is not reliable for J = 3, although the 95% interval coverage is above nominal.

New RCT: implementation effect is causal

- Analysis so far is correlational – perhaps omitted variables drive both implementation and TEs?
- Causal impact of implementation can be studied rigorously because it can be varied systematically.
- We do this in a new RCT with Youth Impact in Botswana where treatment group sessions increased the degree of targeting/tailoring to student abilities.
- Results showed large and significant improvement in learning outcomes relative to the control group which got the standard Youth Impact implementation.

Conclusion: We must evaluate implementation

- Implementation is not a fixed property of an environment but a variable that can be studied and changed.
- In the context of targeted instruction, implementation almost entirely explains heterogeneity in effects across contexts.
- Cross-study evidence is correlational but new RCTs can be run to study these correlations – we provide a blueprint we hope will be adopted for this.
- Without gathering and analysing this information as part of policy evaluation, *we are mostly guessing as to why things succeed*.

Data

Table: Studies Considered for Evidence Aggregation

Authors	State/Country	Treatment Arm	Delivery	Sample Size
Studies included				
Banerjee et al. (2007)	Maharashtra, India	Balshaki Camps	Volunteer	10000
Banerjee et al. (2010)	Uttar Pradesh, India	First UP Camps	Volunteer	9442
Duflo et al (2011)	Kenya	Tracking	Teachers	6000
Banerjee et al. (2017)	Bihar, India	School Volunteers	Volunteer	3325
Banerjee et al. (2017)	Bihar, India	Teacher Camps	Teachers	2474
Banerjee et al. (2017)	Uttar Pradesh, India	UP 10-day Camps	Volunteer	17266
Banerjee et al. (2017)	Uttar Pradesh, India	UP 20-day Camps	Volunteer	13054
Banerjee et al. (2017)	Haryana, India	In-school Teachers	Teachers	11966
Total	5	8	-	73527