SEPTEMBER 2022

CEDIL Syntheses Working Paper 4

The role of implementation in generalisability:

A synthesis of evidence on targeted educational instruction and a new randomised trial

Noam Angrist and Rachael Meager

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL syntheses working papers

The CEDIL Syntheses Working Paper series publishes innovative research syntheses of the impact of interventions in low- and middle-income countries in the form of evidence and gap maps, systematic reviews, meta-analyses and other synthesis products.

About this working paper

Please direct any comments or queries to the corresponding author, Noam Angrist at nangrist@youth-impact.org.

Suggested citation: Angrist, N. and Meager, R. (2022). *The role of implementation in generalisability: A synthesis of evidence on targeted educational instruction and a new randomised trial, CEDIL Syntheses Working Paper 4, CEDIL, Oxford*. Available at https://doi.org/10.51744/CSWP4

This project was funded by the Centre of Excellence for Development Impact and Learning (CEDIL), supported by UK Aid from the UK Government. The views expressed in this working paper do not necessarily reflect the UK Government's official policies.

Cover design: PhilDoesDesign

Photo credit: Youth Impact

The role of implementation in generalisability: A synthesis of evidence on targeted educational instruction and a new randomised trial

Authors

Noam Angrist
University of Oxford, Youth Impact

Rachael Meager
London School of Economics

Acknowledgements

We thank Witold Wiecek for valuable research contributions, as well as Colin Crossley and Patience Derera for excellent grant and programme administration. We thank Tendekai Mukoyi, Thato Letsomo and Moitshepi Matsheng for their programmatic and organisational leadership. Thank you to Stefan Dercon, Clare Leaver, Adrienne Lucas, Eduardo Masset, Jennifer Opare-Kumi, anonymous reviewers, and participants of the University of Oxford and the Centre of Excellence for Development Impact and Learning (CEDIL) research seminar for inputs into earlier drafts of this work. We thank Josh Angrist, CJ Angrist, Abhijit Banerjee, Claire Cullen, Emily Cupito, Andy de Barros, Esther Duflo, Macartan Humphreys, Ashleigh Morrell, and Shobini Mukherji for their useful conversations. We are grateful to the American Economic Association, authors of the original research papers, and the Jameel Poverty Action Lab (J-PAL), which made all data easily accessible and available and advised on various components of the research. We thank Youth Impact, which hosted and administered the grant for this project. Noam Angrist is a co-founder of the non-governmental organisation Youth Impact, which is involved in generating ongoing evidence on the effectiveness of teaching at the right level. The research and implementation were funded by CEDIL, which is supported by the United Kingdom's Foreign, Commonwealth and Development Office.

Contents

List of abbreviations 6

Abstract 7

1. Introduction 8

2. Educational Intervention and Context 14

3. Data for Evidence Aggregation 16

4. Evidence Synthesis 20

4.1 The Bayesian Hierarchical Approach 20

4.2 The Hierarchical Modelling Approach 21

4.3 Meta-regression within the Hierarchical Framework 22

4.4 Assessing Heterogeneity and Generalisability 23

5. Evidence Aggregation Results 26

5.1 Frequentist Random-Effects Results 26

5.2 Bayesian Aggregation Results 29

5.2.1 Basic Hierarchical Model Results 29

5.2.2 Bayesian Meta-regression Results 35

5.2.3 Posterior Predicted Treatment Effects 37

5.2.4 Results with Theory- and Expertise-Informed Priors 38

6. A Model: The Role of Implementation in Identification of Treatment Effects and Generalisability 45

6.1 Defining Implementation 45

6.2 Identification 47

6.3 Bayesian Model and Estimation Performance 50

6.4 Implementation Model Results 54

7. Implementation Can Be Improved: Evidence from A/B Testing in Botswana 56

7.1 Intervention and Study Design 56

7.2 Results 58

8. Conclusion 59

References 61

Appendix A Additional Bayesian Models and Results 64

Appendix B Evidence Aggregation Robustness Checks 71

List of tables, figures, and boxes

Table 1: Studies considered for evidence aggregation 16
Table 2: One-factor implementation model performance in simulations 53
Table 3: Two-factor implementation model performance in simulations....53
Table 4: Implementation model posterior intervals: inference on take-up 54
Table 5: Implementation model posterior intervals: inference on fidelity and take-up 55
Table 6: Botswana sample: learning levels at baseline....57
Table 7: Results of a randomised increase in programme fidelity 58

Figure 1: Active targeted instruction scale-up efforts 15
Figure 2: ASER assessment examples used across 14 countries 17
Figure 3: Frequentist random-effects meta-analysis of ITT effects....27
Figure 4: Frequentist random-effects meta-analysis of TOT effects....28
Figure 5: Bayesian aggregation of all ITT results 30
Figure 6: Bayesian aggregation of all TOT results....32
Figure 7: Bayesian aggregation of ITT split by delivery 33
Figure 8: Bayesian aggregation of TOT split by delivery 34
Figure 9: Summary of Bayesian aggregation results across multiple models 36
Figure 10: Posterior predictive distributions of future effects 37
Figure 11: ITT results under different theory-driven priors 42
Figure 12: TOT results under different theory-driven priors 43
Figure 13: Rubin (1981) model vs joint "mu and tau" model 65
Figure 14: Individual pooling factors from the simple Bayesian model 66
Figure 15: Additional prior robustness checks....66
Figure 16: Mu and tau model with different LKJ priors 68
Figure 17: Full posteriors predictive distributions under different models 69
Figure 18: ITT leave-one-out analysis....70
Figure 19: TOT leave-one-out analysis....72
Figure 20: Distribution of t-statistics in our sample 73

List of abbreviations

ASER	Annual Status of Education Report
CEDIL	Centre of Excellence for Development Impact and Learning
FCDO	UK Foreign Commonwealth and Development Office
ITT	Intention-to-treat
LKJ	Lewandowski-Kurowicka-Joe
LTE	Latent treatment effect
MCMC	Markov chain Monte Carlo
NGO	Non-governmental organisation
RCT	Randomised controlled trial
REML	Restricted maximum likelihood
RMSE	Root mean squared error
RTE	Realised treatment effect
SD	Standard deviation
TaRL	Teaching at the Right Level
TOT	Treatment-on-treated
UNESCO	United Nations Educational, Scientific and Cultural Organization
UP	Uttar Pradesh
USAID	United States Agency for International Development

Abstract

Targeted instruction is one of the most effective educational interventions in low- and middle-income countries, yet the reported impacts of this approach vary, from 0.07 to 0.78 standard deviations (SDs) across contexts. We study this variation and the contextual factors associated with it by combining an evidence aggregation covering 10 study arms with a new randomised trial. The results show that two factors explain most of the heterogeneity in reported effects: the degree of implementation (intention-to-treat or treatment-on-the-treated effects) and the instruction delivery model (teachers or volunteers). Accounting for these implementation factors enables substantial generalisation of effect sizes across contexts. We introduce a new Bayesian model which incorporates implementation information into the evidence aggregation process. The results show that targeted instruction can deliver 0.39 SD improvements in learning on average when taken up, and 0.80 SD gains when implemented with high fidelity, explaining the upper range of effects in the literature. Given the central role of implementation identified in our synthesis, we conduct a new randomised trial to increase programme fidelity in Botswana. The results show additional 0.22 SD gains relative to standard implementation, revealing concrete mechanisms to enhance implementation and achieve the largest frontier effects identified in the literature.

1. Introduction

617 million young people worldwide are in school but are unable to read fluently or to perform simple numerical operations. These learning deficits are particularly acute in developing countries (United Nations Educational, Scientific and Cultural Organization (UNESCO), 2017; World Bank, 2018; Angrist, Djankov, Goldberg and Patrinos, 2021). While many educational interventions have met with limited success, targeting educational instruction to a child's learning level has been shown to improve learning in a variety of contexts.[1] Randomised trials show consistently positive impacts of targeted instruction in India, Kenya and Ghana (Banerjee et al., 2007; Banerjee et al., 2010; Duflo, Dupas and Kremer, 2011; Banerjee *et al*., 2017a; Duflo, Kiessel, and Lucas, 2020).

The consistently positive effects found for targeted instruction have received significant attention in the academic literature and in policy circles.[2] For example, a high-profile report by the World Bank, UNICEF, and the United Kingdom's Foreign, Commonwealth and Development Office (FCDO) highlights targeted instruction as a cost-effective solution to address the global learning crisis (Global Education Evidence Advisory Panel, 2020). However, while the effects are consistently large, they range from 0.07 to 0.78 standard deviations (SDs) – an order of magnitude difference.[3] Systematic analysis of this variation could reveal important information on how generalisable effects might be, as well as factors that can yield the largest 'frontier effects' identified in the literature. This is especially important as targeted instruction is now being adapted across contexts, with multiple ambitious scale-up efforts underway. In this paper, we first assess the generalisability of the evidence on targeted instruction using evidence aggregation. We then use the results to inform randomised optimisations of a targeted instruction scale-up in Botswana. For our aggregation, we consider data across 10 study arms, covering nearly 100,000 students. We collect data on effect sizes, as well as contextual covariates such as baseline learning, geographical context, sample size, year, and instruction delivery model (teachers or an external volunteer). We also consider data on programme implementation first using the notion of 'take-up', measured by attendance or the presence of classroom materials. Secondly, we consider the 'fidelity' of implementation,

[1] Targeted instruction groups students in classrooms and tailors instruction to each student's actual learning level, rather than to an average expected learning level determined by a one-size-fits-all grade-level curricula. A specific model of this approach called 'Teaching at the Right Level' has been pioneered by Pratham, a large education non-governmental organisation (NGO) in India.

[2] Multiple reviews identify targeted instruction as an effective educational approach (Kremer, Brannen and Glennerster, 2013; Snilstveit *et al*., 2016; Angrist et al., 2020).

[3] Of note, these reported effects are all substantial in a context in which a 0.10 standard deviation effect size is considered large (Kraft, 2020; Evans and Yuan, 2020).

measured by the extent to which lessons adhere to core programme principles (i.e. whether instruction is targeted, and students are grouped as expected). Targeted instruction offers an ideal setting in which to study the different roles of these two aspects of implementation, as both vary widely across studies in our dataset: take-up ranges from 8% to 90%, and fidelity from 23% to 83%.

Given heterogeneity in both programme features and reported effects, careful attention to the evidence aggregation methodology is required. We provide results from both the standard frequentist random-effects meta-analysis and a series of Bayesian hierarchical models, including meta-regression models which formally incorporate data on programme features. We also report several metrics of generalisability, including the frequentist I-squared metric, which measures the percentage out of the total variation which is genuine variation in effects (rather than sampling variation), the Bayesian hypervariance, which measures the variation in effects directly, and the posterior predictive distribution, which captures uncertainty about the predicted effect in the next hypothetical study setting. Given a sample of 10 study arms, which is large as regards synthesising evidence but small for frequentist statistical approaches relying on large-sample properties, we prefer the Bayesian approach for evidence aggregation in our setting.[4] Nevertheless, we report both to ensure complete transparency.

Our results show that most of the heterogeneity in effects reported in the literature can be explained by two factors: instruction delivery model (teachers or volunteers) and the degree of implementation (measured via take-up). Conditional on these factors, the effect of targeted instruction is highly generalisable across studies. The frequentist random-effects meta-analysis finds that intention-to-treat (ITT) effects for teachers are moderate (0.07 SDs on average) and highly generalisable (I-squared of 0.01%), whereas volunteers have large average effects (0.22 SDs), with high variation (I-squared of 94.2%). When accounting for implementation, treatment-on-treated (TOT) effects for volunteers are three times larger, with an average effect of 0.76 SDs. Notably, TOT effects are now highly generalisable, with an I-squared of 0.0%. This result reveals that much of the original heterogeneity in ITT estimates was due to variation in implementation.

The Bayesian analysis upholds these patterns, although the results are somewhat tempered.[5] The TOT effect is still much larger than the ITT on average, particularly for

[4] For meta-regression in particular, the priors prevent overfitting by regularising the model fit (Gelman et al., 2004; Hastie, Tibshirani and Friedman, 2009; Meager, 2019).

[5] This is due to accounting for the joint uncertainty on the average effect and the variation in effects (Gelman et al., 2004, Meager, 2019).

volunteer delivery, and highly generalisable across settings. Moreover, individual studies' TOT estimates see large gains in precision due to partial pooling, combining information from low and high implementation settings.[6] Bayesian meta-regression confirms that implementation take-up and instruction delivery model are two key factors in predicting variation in effects. This is not obvious ex-ante, with these two dimensions playing a more substantial role than other factors which a priori could have mattered most, such as students' baseline learning levels. In addition, the evidence on the positive impact of targeted instruction is strong, even when we impose strong priors. This suggests the patterns in the data are robust and informative. We show that these results are robust to dropping any individual study and show no evidence of publication bias. Overall, these results show that features of programme implementation predict the largest effects in the literature, as well as the generalisability of effects across settings.

We introduce a novel Bayesian aggregation model which incorporates implementation information in evidence aggregation.[7] Our model offers a crucial theoretical result: namely, that neither treatment effects nor their variation across settings can be identified in the absence of information on implementation. To give an example, consider the case of null treatment effects. Null effects could be due to an ineffective programme or an effective programme which was never implemented. Without information on implementation, we might erroneously conclude that there is a null treatment effect when in fact it is a case of null implementation.[8] Simulations show that the simplest version of the model performs well. We fit this model to our data considering both programme take-up and programme fidelity as aspects of implementation. Using the available data on both of these variables, our results show that targeted instruction offers 0.39 SD improvements in learning on average when fully taken up, and could deliver up to 0.80 SD gains when implemented with high fidelity, explaining the upper range of effects in the literature.

Our evidence aggregation and model establish the importance of implementation, which calls for conducting research into concrete ways to increase take-up and fidelity. We conduct a randomised trial of an effort to optimise programme fidelity in a targeted instruction

[6] For example, First UP Camps becomes statistically significant by conventional frequentist standards.

[7] In this model, we jointly account for uncertainty on effects and uncertainty in implementation.

[8] A full exposition is provided in the paper. Here we summarise a key component. We define realised treatment effects (RTEs) as equal to latent treatment effects (LTEs), represented by θ_j – that is, treatment effects under full implementation – multiplied by an implementation factor, $m_j \in [0, 1]$ such that $RTE = m_j * \theta_j$. A programme that has no impact could be driven by a situation in which $\theta_j = 0$ but, equally possibly, $m_j = 0$. Without explicit information on m_j, a treatment effect of zero cannot be logically used to infer a null latent treatment effect θ_j. In other words, the underlying effect θ_j cannot be identified from the data.

intervention in Botswana. The Ministry of Basic Education in Botswana, in partnership with Youth Impact, one of the largest NGOs in the country, is in the process of scaling up and continually testing a targeted instruction programme, with over 20% of primary schools in the country reached to date. In the trial, we randomly vary implementation fidelity – achieved via more detailed learning assessments and grouping of students relative to standard implementation – across 52 classes and over 1,000 students in four regions. We find that improved fidelity increases the programme's impact by a precisely estimated 0.22 SDs with baseline controls. These results confirm that the correlation between implementation and impact observed in the literature reflects a causal relationship – it is not merely the case that favourable settings yield both high implementation and large effects, but rather that improving implementation can directly improve programme results in a given setting, holding all else equal. Thus, overall, we find that programme delivery and implementation levels are decisive factors in both the size and generalisability of the programme's impact, and that these factors can be changed in the field to further enhance learning gains as effective programmes scale up.

Our findings contribute to several literatures. First, we contribute to the literature on education in low- and middle-income countries. Improving learning outcomes is difficult, with decades of stagnant learning outcomes, despite increasing enrolment in school (World Bank, 2018; Angrist *et al*., 2021). Moreover, input-focused interventions which simply provide more resources, such as the provision of additional textbooks or computer hardware only, have been found to rarely improve learning (Kremer, Brannen, and Glennerster, 2013). In contrast, pedagogy-focused interventions which aim to improve the quality and type of teaching in the classroom have had far greater success in improving learning, such as targeted instruction and structured pedagogy approaches (Duflo, Dupas and Kremer, 2011; Piper *et al*., 2014; Banerjee *et al*., 2017a; Muralidharan, Singh, and Ganimian, 2019; Duflo *et al*., 2020). Our findings are consistent with this emerging view and show that this insight generalises even across heterogeneous contexts.

Second, we contribute to the literature on external validity and advance the practice of evidence synthesis in economics. Although systematic evidence aggregation is still a relatively recent effort in economics, researchers are increasingly engaging in evidence synthesis across contexts. Vivalt (2020), Banerjee *et al*. (2017b), Bandiera *et al*. (2017), Gechter (2017), Andrews and Oster (2019), and Meager (2019; 2022) have addressed this question for a variety of interventions, with differing kinds of data sources, using both Bayesian and frequentist methodologies. We contribute methodologically to this literature with a new aggregation model incorporating data on implementation, which we hope will be used in other literatures where these data are available. We also contribute new results by showing that targeted instruction has both large and generalisable effects. This stands in contrast to several previous findings showing small or null effects of multiple policies and interventions, such as microcredit (Meager, 2019).

Third, we contribute to a nascent literature on implementation science in education. In particular, we demonstrate the first-order importance of implementation and provide a formal framework to account for the degree and fidelity of implementation in programme evaluation. In the first instance, our results reveal the importance of accounting for implementation by reporting TOT effects. Reporting TOT estimates is a practice that is often inconsistent across disciplines and underutilised in meta-analysis. Out of a set of 400 papers in development from 2019 and 2021, for example, only two of the education randomised controlled trials (RCTs) reported TOT effects.[9] Moreover, reporting TOT effects provides a concrete, quantitative mechanism to account for implementation, in a context where implementation has to date been seen as a 'fuzzy' and qualitative concept, despite a growing recognition of its importance (Bauer *et al*., 2015). Second, our results showing that implementation is the decisive dimension in generalising results across contexts motivates a research agenda focusing on the details of effective implementation, consistent with the notion of 'the economist as plumber' (Duflo, 2017). Third, we formalise the essential role of implementation in the identification of treatment effects and for generalisability. We go a step further and introduce a novel Bayesian evidence aggregation model which incorporates implementation. We find the average effect of targeted instruction is 0.39 SDs when taken up, and 0.47 SDs for volunteer-delivered programmes. These effect sizes are up to fivefold higher than the typical effective education intervention, which on average improves learning by 0.10 SDs. These results suggest that research into increased uptake of productive interventions, such as targeted instruction, can offer a much higher return on investment than discovering new effective interventions.

The results in this paper have significant implications for policy. Targeted instruction has been featured in multilateral policies and lists of evidence-based interventions as a high-potential approach to address the global learning crisis, including by the World Bank, FCDO,

[9] We conduct a review to identify how frequent the practice of accounting for implementation is in programme evaluation. Out of a set of 400 papers in development from 2019 and 2021 in a set of top economics journals, 19% were RCTs; of those that were RCTs in education, only two papers reported TOT estimates. The set of journals considered includes the top five economic journals (American Economic Review, Quarterly Journal of Economics, Econometrica, Journal of Political Economy, and Review of Economic Studies) and other top-tier general interest journals (Review of Economics and Statistics, Economic Journal, Journal of the European Economic Association, and all four American Economic Journal journals), and a top field journal (the Journal of Development Economics).

and the United States Agency for International Development (USAID). It is also on track to reach over 60 million children cumulatively in South Asia and sub-Saharan Africa by 2025 (J-PAL, 2022).[10] High-profile scale-up examples include Zambia, where the government has already scaled up targeting instruction to over 2,000 schools, and Nigeria, where targeted instruction is being delivered in over five states. In addition, in Botswana the government has signed a nine-year memorandum of understanding with Youth Impact, one of the largest NGOs in the country, relating to scaling up targeted instruction nationally. The results in this paper inform the generalizability of targeted instruction, as well as contextual factors and programme components that mediate the largest effects identified in the literature, and thus provide policy-relevant insights as these programmes are adapted and scaled up in new contexts.

The remainder of the paper is organised as follows. Section 2 describes the intervention and its context. Section 3 describes the data used for the aggregation exercise, and Section 4 describes the Bayesian hierarchical modelling approach. The results of the aggregation exercise are presented in Section 5. In Section 6 we propose a new Bayesian synthesis model which incorporates implementation in the evidence aggregation process, including both take-up and fidelity. This includes formalising the role of implementation in the identification of treatment effects and for generalisability, model performance assessments via simulations, and results. Section 7 includes results from a new randomised trial in Botswana to optimise targeted instruction fidelity as the programme scales up, and Section 8 concludes, providing ideas for further research.

[10] https://www.povertyactionlab.org/case-study/teaching-right-level-improve-learning

2. Educational Intervention and Context

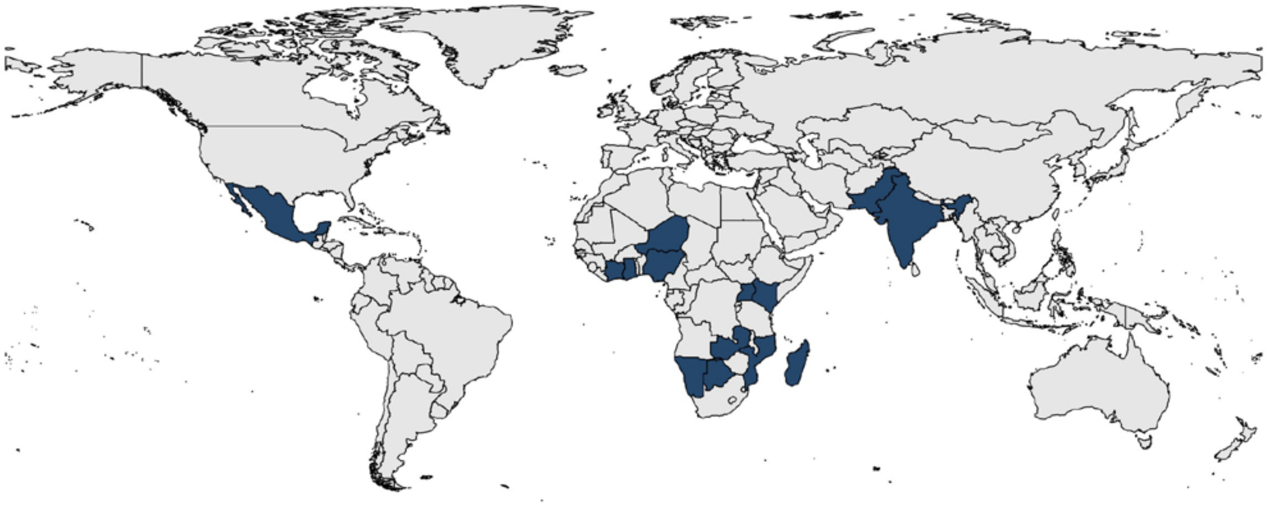
Educational enrolment has increased worldwide over the last few decades to above 90% in all regions of the world, yet learning progress has remained limited. UNESCO and the World Bank have called this phenomenon a 'learning crisis' (World Bank, 2018). This learning crisis is most pronounced in low- and middle-income countries. For example, in Kenya, Tanzania and Uganda three-quarters of Grade 3 students cannot read a basic sentence such as 'the name of the dog is Puppy'. In rural India, half of Grade 3 students cannot solve a two-digit subtraction problem, such as 46 minus 17 (World Bank, 2018). The global learning crisis is estimated to cost over US$ 129 billion in lost social welfare (UNESCO, 2017).

A combination of factors contribute to the learning crisis, including curricula targeted mostly to advanced students, rote learning, and automatic promotion regardless of the learning achieved in prior grades (Banerji and Chavan, 2016). Many education interventions have focused on providing inputs to improve learning, such as textbooks, computers or cash transfers, reducing class sizes, or increasing teacher salaries. However, decades of randomised trials across countries show input-focused initiatives rarely improve learning outcomes (Kremer, Brannen, and Glennerster, 2013; Evans and Popova, 2016; Ganimian and Murnane, 2016; Snilstveit *et al*., 2016; Angrist *et al*., 2020).

In contrast, a pedagogical shift – targeting instruction to the level of the child – has been shown in randomised trials to dramatically improve learning across multiple con texts, including India, Kenya (Kremer, Brannen, and Glennerster, 2013) and Ghana (Duflo *et al*., 2020). Targeted instruction involves regrouping students by their actual learning level (e.g. in relation to addition, subtraction), rather than using grade-level grouping only, which teaches to the expected learning level, which is determined by often over-ambitious curricula. Most education systems are organised to teach a one-size-fits-all curriculum by grade. However, there is often substantial heterogeneity in student learning levels in each grade, with most students well below grade-level expectations. Targeting instruction involves regrouping students across grades and instead teaching students based on their by proficiency levels. Instead of using mass education that reaches only a few, this approach uses customised and engaging teaching and learning that is targeted to the level of every child. For example, a teacher's syllabus might prescribe them to teach division to a class of Grade 3 students, which is the curriculum-level expectation. However, if only 10% of the class understands division, 90% of the class will be left behind. If a child cannot recognise or add numbers, they will not be able to learn division. Indeed, data from India show that most children are far behind the expected grade level. The Annual Status of Education Report (ASER), a nationally representative survey or rural areas in India, showed that only 51% of Grade 3 students could read a Grade 1 text (ASER, 2019). This phenomenon has also been documented extensively in sub-Saharan Africa.

Active targeted instruction scale-up efforts are ongoing in Botswana, Cote d'Ivoire, Ghana, India, Kenya, Madagascar, Mexico, Mozambique, Niger, Nigeria, Pakistan, Uganda, and Zambia. For more information see https://www.teachingattherightlevel.org/).

Figure 1: Active targeted instruction scale-up efforts



A specific model for this intervention called 'Teaching at the Right Level' (TaRL) has been developed by Pratham, one of the largest education NGOs in India. TaRL specifically, and targeted instruction more broadly, has been shown to consistently improve learning outcomes for children across diverse contexts.[11] However, as yet there is no systematic assessment using meta-analysis across studies to identify factors which drive heterogeneity and predict the highest frontier effects in the literature. Making progress on this question would have significant policy implications as the targeted instruction approach is actively being adopted by dozens of countries and scaled up to over 60 million children worldwide. The World Bank, USAID, FCDO, governments, and NGOs are all engaged in large-scale targeted instruction scale-up efforts. Figure 1 highlights active scale-up efforts in Botswana, Cote D'Ivoire, Ghana, India, Kenya, Madagascar, Mexico, Mozambique, Niger, Nigeria, Pakistan, Uganda and Zambia.

[11] Both in-school as well as some edtech models of targeting instruction have been showing to be effective. In this paper we focus largely on in-school and Pratham-delivered models. However, other targeted instruction models exist, such as mindspark software, which adapts to the level of the child (Muralidharan, Singh, and Ganimian, 2019), as well as low-tech phone-based tutorials (Angrist, Bergman, and Matsheng, 2022).

3. Data for Evidence Aggregation

Studies included. We analyse microdata across a series of existing clustered RCTs conducted over the last two decades across India and Kenya. In a future version of this paper, we will also incorporate upcoming results from studies in Ghana. In total, these trials represent 14 geography-treatment arm trial units, and 10 study arms (not including Ghana). The total sample across the studies is nearly 100,000 students. We start our analysis with these studies since they are recognised as consistent with the targeted instruction model in developing contexts by both original evaluators and implementors. Moreover, we can access the microdata for all these studies, which allows us to replicate the original results as well as enhance our ability to conduct a comprehensive evidence aggregation exercise capturing study-level covariates and programme features. We focus our analysis on RCTs to ensure that we aggregate causal effects.

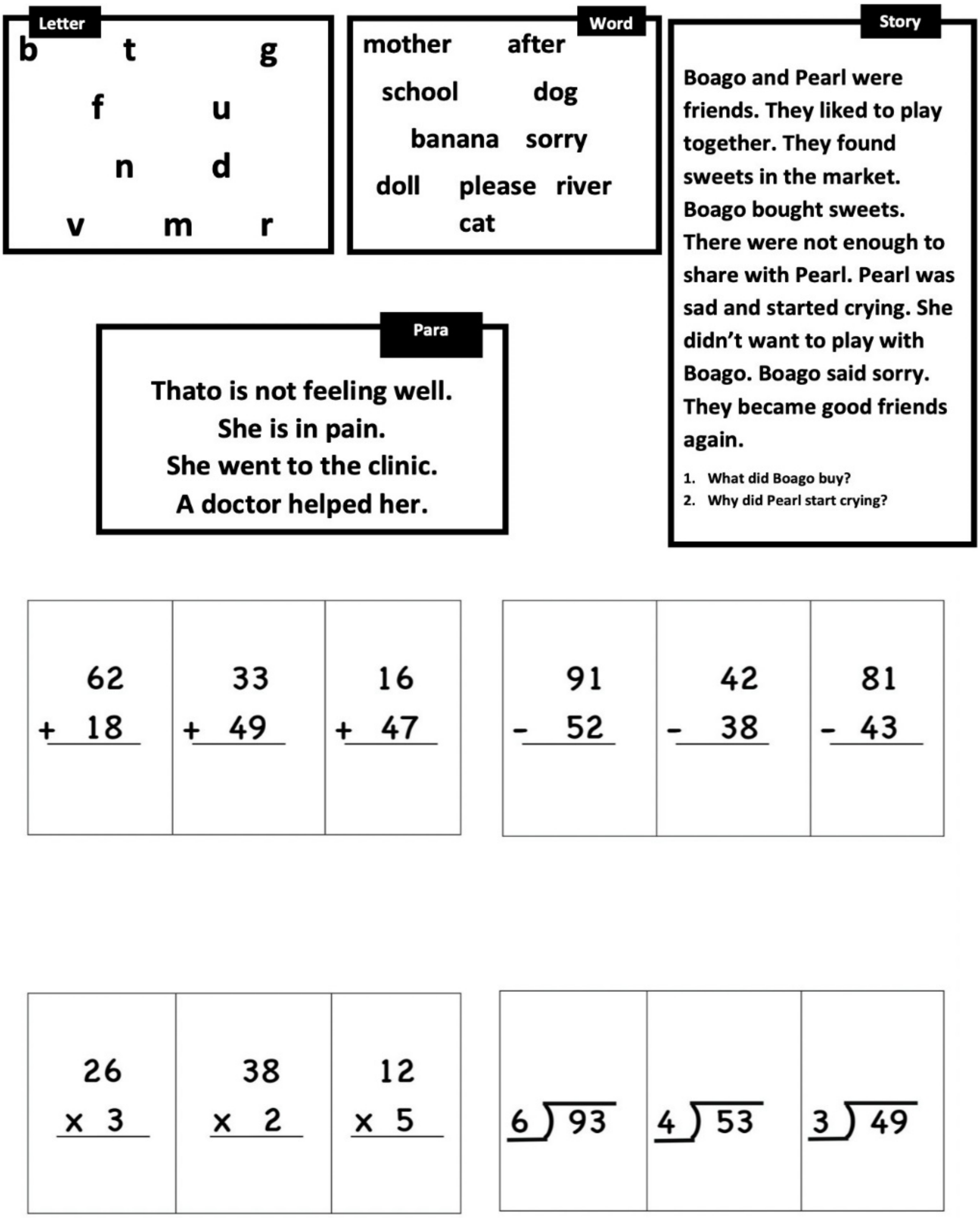
Table 1 lists the programme studies considered and highlights key sample characteristics for each study. We consider the relevant level of observation to be the study-treatment arm-geography. States in India are highly heterogeneous and it would be potentially misleading to categorise them together as a single country. We include states in India as geographical units; moreover trials in India often stratify or randomise within a state (as well as across states).

Table 1: Studies considered for evidence aggregation

Authors	State/country	Treatment arm	Delivery	Sample Size
Studies included				
Banerjee *et al*. (2007)	Maharashtra, India	Balshaki Camps	Volunteer	10,000
Banerjee *et al*. (2007)	Gujarat, India	Balshaki Camps	Volunteer	11,000
Banerjee *et al*. (2007)	Gujarat, India	Computer-assisted	Volunteer	6,131
Banerjee *et al*. (2010)	Uttar Pradesh, India	First Uttar Pradesh (UP) Camps	Volunteer	9,442
Duflo *et al*. (2011)	Kenya	Tracking	Teachers	6,000
Banerjee *et al*. (2017a)	Bihar, India	School Volunteers	Volunteer	3,325
Banerjee *et al*. (2017a)	Bihar, India	Teacher Camps	Teachers	2,474
Banerjee *et al*. (2017a)	Uttar Pradesh, India	UP 10-day Camps	Volunteer	17,266
Banerjee *et al*. (2017a)	Uttar Pradesh, India	UP 20-day Camps	Volunteer	13,054
Banerjee *et al*. (2017a)	Haryana, India	In-school Teachers	Teachers	11,966
Studies to be included				
Duflo *et al*. (2020)	Ghana	In-school Remedial	Volunteer	2,000
Duflo *et al*. (2020)	Ghana	After-school Remedial	Volunteer	2,000
Duflo *et al*. (2020)	Ghana	In-school not Targeted	Teachers	2,000
Duflo *et al*. (2020)	Ghana	In-school Targeted	Teachers	2,000
Total	**8**	**14**	**-**	**98,658**

Outcome Data and Measurement. We assess the microdata from each study to produce standardised outcome variables. In all studies, the central outcome is a measure of learning on basic numeracy and literacy skills. Most studies use an assessment similar to the ASER test. Figure 2 shows examples of ASER assessments for literacy and numeracy.

Figure 2: ASER assessment examples used across 14 countries



Letter

b t g
f u
n d
v m r

Word

mother after
school dog
banana sorry
doll please river
cat

Story

Boago and Pearl were friends. They liked to play together. They found sweets in the market. Boago bought sweets. There were not enough to share with Pearl. Pearl was sad and started crying. She didn't want to play with Boago. Boago said sorry. They became good friends again.

1. What did Boago buy?
2. Why did Pearl start crying?

Para

Thato is not feeling well.
She is in pain.
She went to the clinic.
A doctor helped her.

62 + 18	33 + 49	16 + 47

91 - 52	42 - 38	81 - 43

26 x 3	38 x 2	12 x 5

6) 93	4) 53	3) 49

The ASER test is a validated learning measure which tests competencies used across 14 countries and it is consistently used in the education literature. In numeracy, questions include number recognition, addition, subtraction, multiplication, and division. In literacy, the competencies tested include letter recognition, word recognition, ability to read a sentence fluently, and reading comprehension of a paragraph and short story. The Kenya study is the only study which uses a different assessment, which is a 100-point test which also covers basic numeracy and literacy. We examine average scores over both numeracy and literacy. In most cases both subjects are available; however, in two cases only one subject is available. All studies include baseline and endline data and some studies also include midline data. For consistency, and to capture longer-lasting effects, we focus on effects at endline. To compare these outcomes, we standardise each score relative to the standard deviation within a state/country-treatment arm unit. We derive learning gains over the course of one year to compare outcomes on a consistent time horizon. Given that our underlying assessments measure a similar outcome and most use a similar test we are relatively confident in the comparability of outcomes across contexts.

One of the advantages of the targeted instruction intervention is that the outcomes are similar across intervention settings. In many cases, meta-analyses in economics rely on outcomes which can vary substantially, derived using entirely different surveys and even different definitions of outcome variables, as may be demanded by the different contexts, such as in the case of microcredit (among others) (Meager, 2019; Vivalt, 2020; Pritchett and Sandefur, 2015). Our relatively uniform outcome data are particularly well-suited to aggregation and offer a substantial improvement over many meta-analyses in economics in regard to the comparability of treatment effect across contexts. Throughout the paper we use control group SD units, as this is the usual format of results in the education literature.[12]

Implementation data. Most studies in our sample report ITT effects using randomisation assignment to estimate treatment effects. We replicate these along with the TOT effects for those who actually received the programme. In some cases TOT effects were originally reported, but not in all cases, so we calculate new estimates for all studies, for consistency. Capturing the degree of implementation via the take-up variable is likely to be central to understanding both the average impact and the generalisability of the evidence, since in some studies the degree of implementation is over 80%, while in other studies it is around 10%. It is further noteworthy that in many cases implementation is very high. This reveals that while

[12] The practice of using SD units is widespread, but not often formally justified. Comparable effects on the raw scale may diverge in regard to SD units, or vice versa; this is an important topic for future work.

implementation can vary, it can also reach near-complete levels, increasing the relevance of understanding effects under full take-up.

We assess the microdata from original studies to quantify the degree of implementation. We have take-up data for nearly all studies; only two treatment arms do not record this information. As a result, when we present TOT results, we have two fewer treatment arms compared to ITT results.

More broadly, we consider four measures to capture distinct aspects of programme implementation: (a) teacher attendance, (b) student attendance, (c) materials usage, and (d) whether students were grouped by learning level. The last measure is ideal since it most directly captures fidelity to the intervention; however, this measure is only available in three intervention-arms thus far, all of which are in Uttar Pradesh. We incorporate this information where possible throughout our analysis, with appropriate caveats due to the small sample size.

Additional data and covariates. We standardise and incorporate a series of additional data and covariates that are likely to mediate effects across studies. These include the following: geography (country or state), delivery model (teacher or volunteer), year of intervention, baseline learning levels, and the degree of implementation. A series of additional model dimensions might be important, such as whether the intervention was conducted during school hours or after school hours, and could merit further future exploration.

Replication. We replicate original results prior to conducting a new meta-analysis. By and large, we find consistent results with original reported estimates. Average differences across all studies are less than 0.1 SDs. In a few rare cases, the magnitude of estimates differs slightly from the original estimates. Reasons for this variation include the time horizon of the original reported estimates, the use of a midline rather than endline assessment, different constructions of SDs, and our primary measure being an average of scores across both subjects, that is, numeracy and literacy. In addition, in our replication we do not include control variables that can have minor effects on final estimates.

4. Evidence Synthesis

We systematically analyse the variation in treatment effects of targeted instruction. We conduct various types of meta-analyses. An advantage of meta-analyses is that they are a systematic and quantitative method that can be used to synthesise evidence. Moreover, they generate statistics which can be used to gauge average effect sizes across contexts and so can be used to assess generalisability. In some disciplines meta-analyses are seen as the next tier in evidence strength after RCTs, enabling systematic aggregation of internally valid studies across studies and contexts. However, others argue that meta-analysis are atheoretical and often compute average effect sizes without creating coherent classes of interventions to aggregate (e.g. averaging effects of interventions that are quite different).

In this paper, we aim to draw on the benefits of systematic and quantitative study aggregation, while also ensuring we aggregate coherent classes of interventions and delivery models, and informing our meta-analysis by drawing on theory and qualitative expertise. Bayesian synthesis lends itself particularly well to this approach, especially since it makes it possible to capture expertise and theoretical insight through informed choices of priors. We use both frequentist random-effects synthesis, which is a typical meta-analysis approach, and Bayesian hierarchical models, which is our preferred approach and which we outline below.

4.1 The Bayesian Hierarchical Approach

Aggregating evidence from different settings in the presence of concerns about generalisability requires joint estimation of both average effects and heterogeneity in effects across studies. The statistical challenge is to separate genuine heterogeneity in effects from sampling variation and simultaneously to use this variation to inform the uncertainty about the average impact. Hierarchical models are able to perform this decomposition (Gelman *et al*., 2004; Meager, 2019). However, the interdependent uncertainty between the means and the variances creates a potentially challenging joint inference problem, particularly with a small number of studies. In this setting, Bayesian methods can offer improved tractability and estimation performance relative to popular frequentist counterparts, such as random effects or empirical Bayes (Rubin, 1981; Gelman *et al*., 2004; Gelman and Hill, 2007; Chung *et al*., 2013; Chung *et al*., 2015). The Bayesian hierarchical framework also permits multiple comparisons, automatically adjusting for multiple testing problems, because marginalisation of the joint posterior appropriately conditions on all the evidence available in the sample and priors.

We use a set of Bayesian hierarchical models which estimate the average treatment effect across all studies and the variance across contexts, in line with Rubin (1981), Gelman *et al*. (2004), Vivalt (2020), Bandiera *et al*. (2017) and the Cochrane Handbook version 5.1 section 16.8. This approach provides an initial estimate of the degree of generalisability in these average effects, without conditioning on any specific information about the individuals involved or the programmes being implemented. In this section we discuss the details of this modelling approach (subsection 4.2), with particular attention given first to meta-regression (subsection 4.2) and then to measuring generalisability (subsection 4.4), due to our interest in the role of programme features in predicting effect size and heterogeneity across settings.

4.2 The Hierarchical Modelling Approach

We take as our foundational approach the canonical Rubin (1981) 'Eight Schools' Bayesian hierarchical model. This model has been extensively used in the literature and considers a set of J total estimated treatment effects $\hat{\theta}_j$ and their standard errors $\hat{se}_j$ (Rubin, 1981; Gelman *et al*., 2004; Meager, 2019). The estimates are typically assumed by authors of empirical research papers to be normally distributed around the true effects θ_j , by their using a consistent estimator and invoking the central limit theorem. These assumptions underlie the computation of confidence intervals and p-values in frequentist research papers, and this therefore imposes no additional structure over the original papers' analyses. The hierarchical component of the model then additionally posits that these effects are normally distributed around some true average or 'hypermean' effect θ, with some 'hyper standard deviation' or 'hyperSD' σ_θ governing their dispersion around the true effect. The resulting hierarchical likelihood is often written as follows:

$$\begin{aligned} \hat{\theta}_j &\sim N(\theta_j , \hat{se}_j) \\ \theta_j &\sim N(\theta , \sigma_\theta^2) \end{aligned} \qquad \text{(Equation 4.1)}$$

As discussed in Meager (2019), this parametric model is more general than it appears: if the hyperSD is set to 0 this model nests classical frequentist fixed-effects meta-analysis, while if the hyperSD is very large the model nests the no-pooling case, in which no aggregation is possible due to heterogeneity across studies. The model can be estimated in a frequentist manner, although there are often computational difficulties in practice (see Meager (2019) for a complete discussion). It can be easier for this model to be estimated using Bayesian methods, either with improper priors or with priors informed by statistical or economic reasoning. The model's performance has been extensively discussed in Gelman *et al*. (2004) and it is generally known to have good frequentist properties, including attaining nominal coverage rates for the posterior credible intervals: i.e. the central 95% posterior intervals contain the true parameter 95% of the time. The key assumption embedded in the model is that of exchangeability between the effects being studied, which is a weaker form of the classical independent and identically distributed (i.i.d). assumption, and usually reasonable in meta-analytic settings (see Meager, 2019 for more discussion).

To estimate this likelihood model in the fully Bayesian manner, and thus derive the joint posterior, it is necessary to add prior distributions to the hyperparameters. Following Gelman *et al*. (2004) and Meager (2019), in this paper we use weakly informative priors as a default approach: this imposes a gentle structure on the problem without unduly influencing the posterior results. For the hypermean, we tend to centre our prior at zero, with a wide uncertainty interval, reflecting the principle that researchers ought to have as their 'null hypothesis' the contention that an untested intervention or policy should be considered most likely to have no impact until proven otherwise by the data. For the hypervariance, we use half-normal or half-Cauchy priors, as suggested in Gelman and Hill (2007), which allows for large variation in effects across settings. As discussed in Meager (2019), and further below in the meta-regression models section of this paper, the priors can improve overall estimation by making a favourable bias-variance tradeoff even when the prior information is incorrect. It is of course possible to incorporate more substantive information, perhaps based on theoretical or contextual knowledge, into the priors. As this is not yet standard practice, we use default priors for our main analysis and build in priors informed by theory. As an additional precaution, we also show prior robustness checks throughout our analysis.

4.3 Meta-regression within the Hierarchical Framework

As our goal is to understand the factors that influence the effectiveness of targeted instruction, we use meta-regression to explore the role of programme-level covariates. Meta-regression is straightforward within the Bayesian hierarchical approach: in the Rubin (1981) model, for example, one need only replace the hypermean θ with a conditional hypermean expression in the style of linear regression. As discussed in Meager (2019), this can be implemented using the following model. Given a set of K contextual factors and covariates, defined by the K-length vector X_j for site j, one can specify a K-dimensional parameter β, such that the expected value of the effects θ_j is the conventional regression surface, and hence E [θ_j] = $X_j\,\beta$. This is implemented via the hierarchical meta-regression likelihood below:

$$\hat{\theta}_j \sim N(\theta_j\,,\,\hat{se}_j)$$

(Equation 4.2)

$$\theta_j \sim N(X_j\,\beta,\,\sigma_\theta^2)$$

While, again, the model is fully parametric, it is more general than it appears. If one were to discard information about sampling variation and assume *theta*$_j = \hat{\theta}_j$ then the model above corresponds in expectation to classical frequentist meta-regression as it is usually conducted in the literature. This is because, as noted in Meager (2019) and elsewhere, the kernel of the Gaussian likelihood corresponds to the ordinary least squares objective function. Of course, in the context of the hierarchical model, it is possible to see a variety of different results due to the partial pooling of information across studies, and thus improved estimation of θ_j during the process of meta-regression itself.

One important note about this model is that in a context with few studies, say J < 15, and many covariates of interest, say K > 3, the estimation of the regression coefficients β becomes very challenging. The primary issue at this scale is the variance of the estimator for the coefficients, which is likely to be large: in short, the estimation is noisy. Moreover, the problem may be masked due to the high risk of overfitting in such a context, leading to false confidence in what we can learn at the study level. However, this issue can be addressed via the use of a machine learning technique known as regularisation (Hastie *et al*., 2009). Regularisation in general involves the incorporation of a penalty function to prevent an estimation procedure from freely wandering around the parameter space. Classical examples include ridge regression, which imposes a squared penalty on the size of the estimated regression coefficients, and lasso, which imposes an absolute value penalty on the same quantity. Within the Bayesian context, it is natural to use the priors to impose the penalty. In fact, as discussed in Hastie *et al*. (2009), in Bayesian analysis a Gaussian prior on the regression coefficients centred at zero is analytically identical to a frequentist ridge regression penalty. We therefore use this penalty throughout our hierarchical meta-regression.

4.4 Assessing Heterogeneity and Generalisability

Within the Bayesian framework, there are several approaches to assessing heterogeneity in effects, and thus the generalisability of the results. The hyperSD or hypervariance parameters in each model capture the population variation in effects in a single parameter, and thus deserve particular attention and reporting. Throughout the paper we report the results on the hyperSD, which is usually treated as the fundamental parameter in Bayesian hierarchical models (Gelman *et al*., 2004). However, it is challenging at times to know how large or small a particular hypervariance estimate is, or how best to interpret it. Thus, we provide two additional metrics of heterogeneity: the frequentist I-squared metric, and the Bayesian posterior predictive distribution.

The I-squared metric measures the percentage of the total variation in estimated effects around the hypermean that is due to genuine variation in true effects, rather than due to sampling variation that causes estimates to vary more than the true effects. This is the metric that is the reciprocal ofthe conventional Bayesian pooling factor discussed in Meager (2019) and Gelman and Pardoe (2006), which measures the percentage of total variation in effects attributable to within-study sampling variation. When I-squared is high, and thus pooling is low, it indicates that the heterogeneity across settings dominates the uncertainty within settings in the sample of studies available. This makes extrapolation across settings challenging and suggests low generalisability of the present results to other settings. Conversely, when I-squared is low and pooling factors are high, the uncertainty within studies about the true effects is larger than the heterogeneity across settings, and thus larger than any hypothetical extrapolation error in the present sample. This corresponds to relatively high external validity, and the Bayesian hierarchical model will perform more 'partial pooling' in such a case, driven by the small estimated hyperSD. In the case of meta-regression, these parameters capture the residual heterogeneity or generalizability, rather than unconditional heterogeneity, and interpretation should be adjusted in that case; we focus on the unconditional hyperSDs and I-squareds for ease of interpretation.

Posterior predictive distributions provide another metric by which we can assess the importance of the heterogeneity in effects across settings. These distributions capture the uncertainty about the hypothetical treatment effect in the next targeted instruction study, given the posterior uncertainty we have about the hypermean and hypervariance. For example, in the Rubin (1981) model, if we define the posterior distribution of these hyperparameters to be $F(\theta, \sigma_{\theta})$ then the posterior predictive distribution for the next effects is:

$$\theta_{J+1} \sim N(\theta, \sigma_{\theta}^{2} \mid F(\theta, \sigma_{\theta}^{2}). \qquad \text{(Equation 4.3)}$$

If one uses an aggregation approach that does not explicitly measure heterogeneity in effects across programmes or studies, such as a fixed-effects meta-analytic model, the posterior predictive distribution is simply the posterior distribution of the hypermean itself. This is because there is no specified heterogeneity in effects across studies in such a model, and thus no quantification of the cross-study extrapolation error. Hence, the extent to which the Bayesian hierarchical posterior predictive distribution is wider than the posterior distribution on the hypermean indicates the extent of heterogeneity in the effects – and, more importantly, it captures the way in which that heterogeneity across settings impedes our ability to extrapolate evidence to the next targeted instruction intervention. This is a natural metric of generalisability of the present evidence base to future settings, and, as such, we report it throughout.

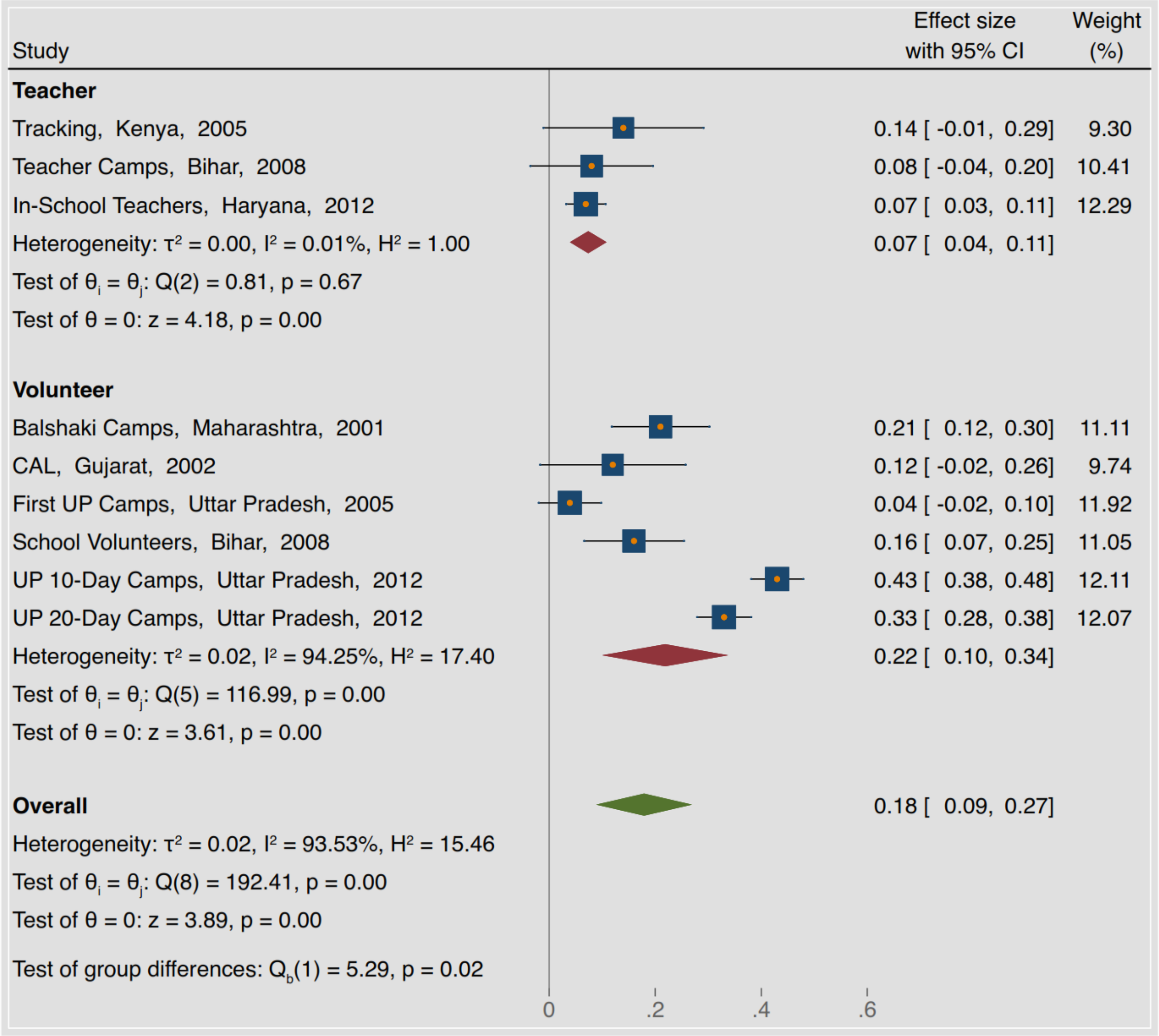
5. Evidence Aggregation Results

5.1 Frequentist Random-Effects Results

In order to contextualise the Bayesian aggregation results, and to provide results without any formal incorporation of theory or priors, we conduct a frequentist random-effects aggregation using the meta-analytic functions in Stata.

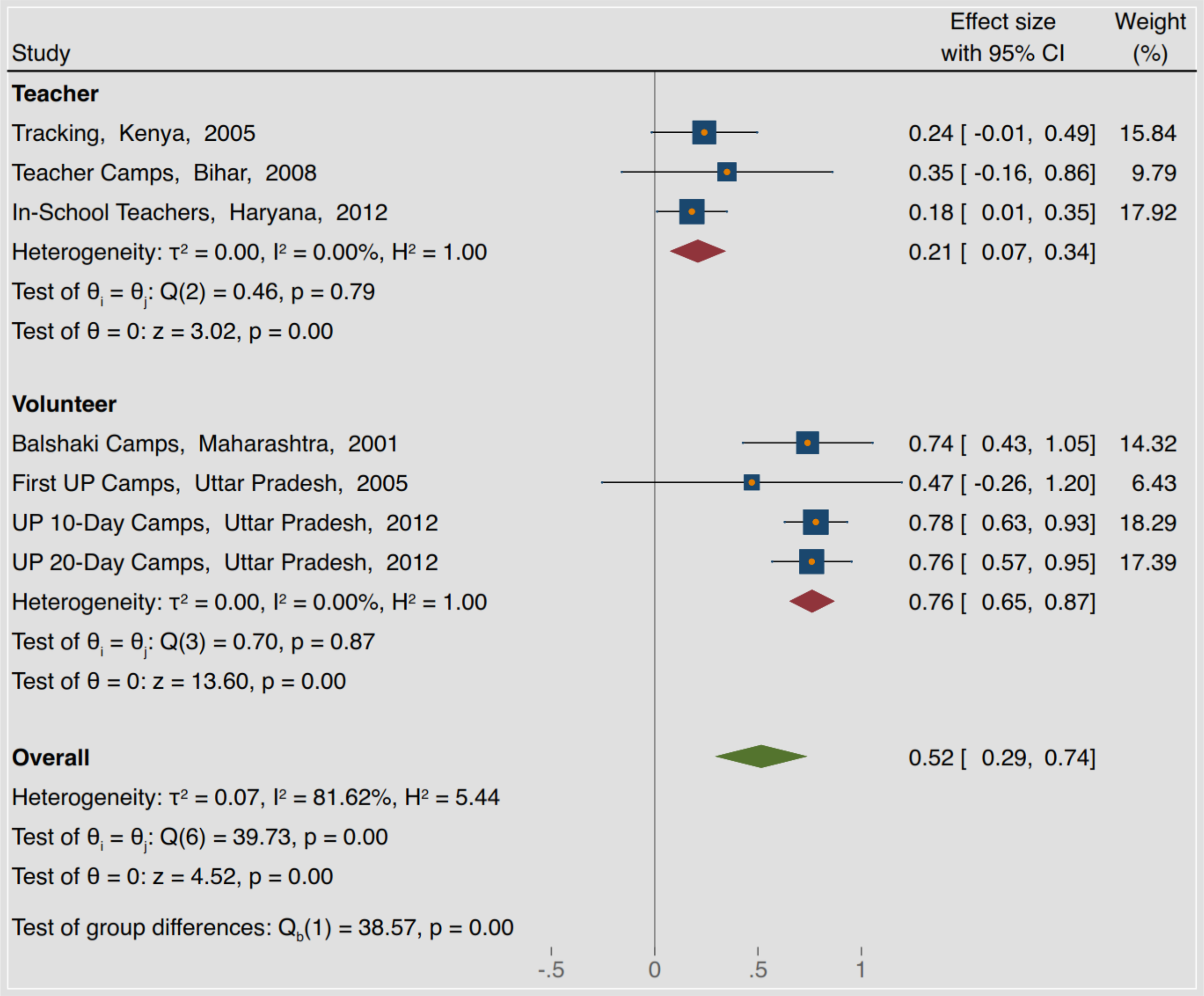
First, we aggregate the evidence on the ITT effects, shown in Figure 3. We find that ITT effects for interventions delivered by teachers have an average effect of 0.07 SDs. These effects are consistent with an I-squared of zero, suggesting any variation between estimates is sampling variation rather than true heterogeneity. This implies the teacher delivery of targeted instruction is extremely generalisable across the programmes in our dataset. Second, we observe that volunteer delivery is on average three times as effective as teacher delivery, with a 0.22 SD effect. However, the volunteer results are highly heterogeneous, with an I-squared of 94.2%.

Second, we aggregate the evidence on the TOT effects that we constructed from the microdata in each study, where available. The results are shown in Figure 4. We observe two trends. First, both teachers and volunteers are three times more effective when implementation is taken into account, with 0.21 and 0.76 SD average effects, respectively. Both effects are large and precisely estimated. Moreover, we now observe convergence among volunteer effects, now with an I-squared of 0%. This suggests that much of the heterogeneity in the volunteer ITT estimates was due to variation in implementation.



Random-effects REML model

Figure 3: Frequentist random-effects meta-analysis of ITT effects



Random-effects REML model

Figure 4: Frequentist random-effects meta-analysis of TOT effects

These preliminary results reveal the initial degree of generalisability, which is high conditional on delivery model and implementation level. These features are also likely mediators of the lowest versus largest effects identified in the literature, and the high generalisability after conditioning on them leaves little room for a role for other features, such as baseline learning levels, although this is only suggestive and will be more formally explored in Section 4.2. It is also worth noting that the patterns in this analysis do not support the notion of diminishing returns relative to programme scale: some of the largest effects, such as those in the Uttar Pradesh 10- and 20-day camps, have the largest sample size (with up to 17,000 students).

5.2 Bayesian Aggregation Results

We now present the results of the Bayesian evidence synthesis and aggregation exercise. First, the results of basic models fit to both the ITT and TOT results. These models correspond conceptually to the frequentist random-effects model, but with joint estimation of the variance in effects and the mean over all the studies, rather than sequentially (e.g. partial pooling), and with the potential to incorporate various choices of priors informed by qualitative expertise, the literature, and theory.

5.2.1 Basic Hierarchical Model Results

We fit the basic Rubin (1981) model to the ITT and TOT estimates from the targeted instruction studies. This model corresponds to the frequentist random-effects model but with joint estimation of the variance in effects and the mean over all the studies, rather than sequentially, and with wide priors centred at zero, to somewhat – though not substantially – discipline the estimation, given the small number of studies. We compare results using partial pooling in our Bayesian aggregation directly to the no pooling case to understand the extent of information pooling across contexts.

Figure 5 displays the results of fitting the basic hierarchical model to the ITT effects of all studies, and Figure 6 shows the results for all available TOT effects. The broad patterns found in the frequentist analysis are confirmed in these two figures: the ITT is much smaller than the TOT on average, and also more heterogeneous. However, there are several interesting differences to note. ITT estimates are relatively unchanged when pooled using Bayesian aggregation: there is slightly more pooling but it remains negligible overall. This is due both to the relative precision of the ITT estimates and to their heterogeneity across settings. Further confirming this, in Appendix A A.1 and A.2 (Figure 14 and Figure 15), we report Bayesian pooling factors, which are the reciprocals of the I-squared metric. The ITT sees only 10% pooling, which is to say that overall only 10% of the variation in results is attributable to sampling variation.

However, the Bayesian model partially pools the TOT estimates together to a substantial degree, due to their greater within-study uncertainty and lesser across-study uncertainty. (See Appendix A (Figure 14) for a confirmation of 25% pooling on average in the TOT results.) We thus observe that the precision of each study’s TOT estimate is enhanced significantly. For example, ‘First UP Camps’ effect, which are positive but not significant in the no-pooling case, are now statistically significant under partial pooling. This is likely due to the fact that implementation was low in this setting (only 8% of students attended sessions) so TOT effects are hard to estimate and are inherently noisy. Pooling studies with high implementation more precisely capture information about the latent effect under full implementation, when average TOT effects are relatively homogenous across studies, as is true in our case, Bayesian aggregation pools TOT estimates where implementation is low with TOT estimates with high implementation studies, substantially enhancing precision.

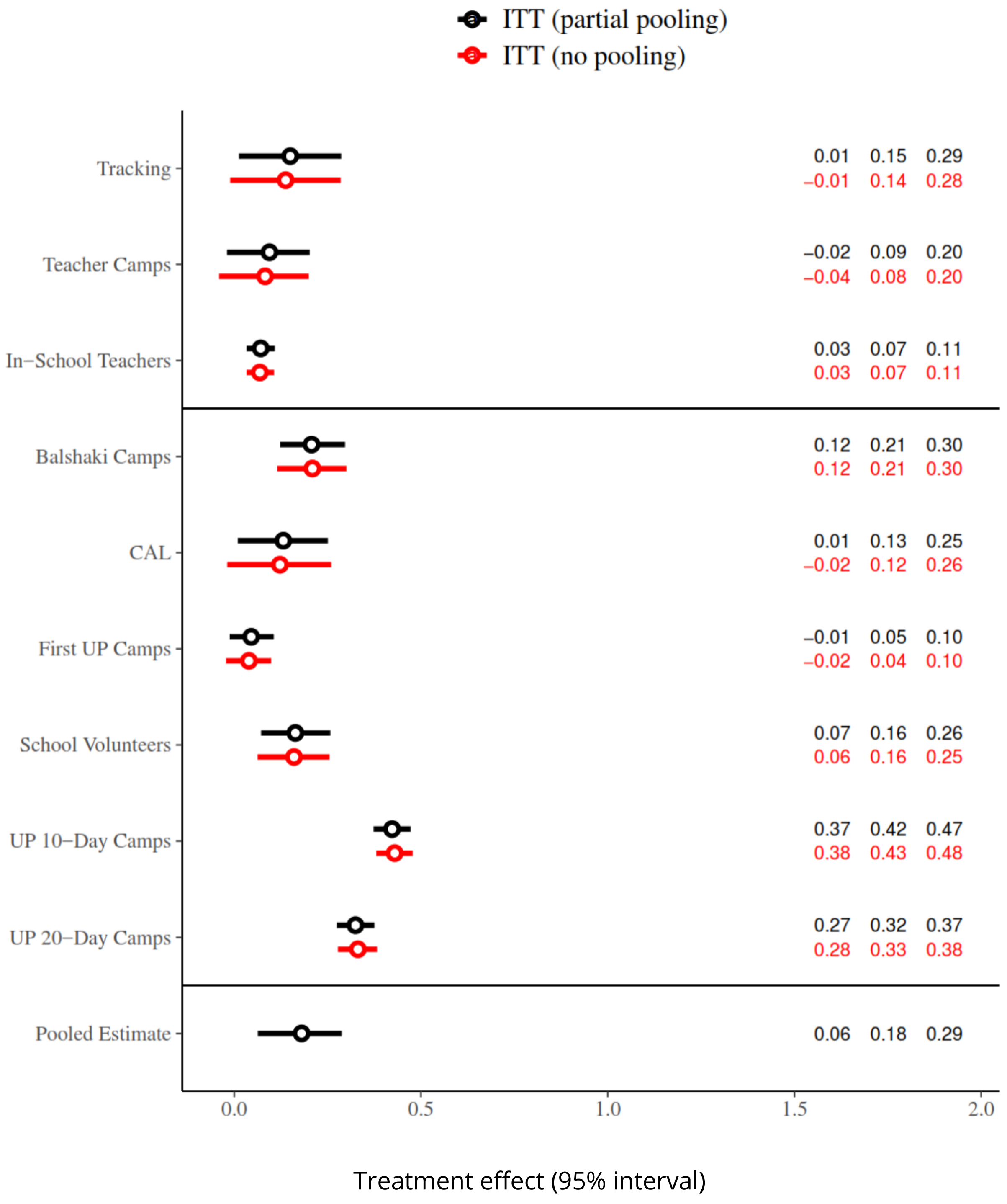


Figure 5: Bayesian aggregation of all ITT results

A central takeaway is that both the average ITT effects and the TOT effects are large and positive. A secondary takeaway is that the TOT effects are three times as large as the ITT effects. Even accounting for the joint uncertainty and using priors that somewhat regularise results towards zero under higher uncertainty, both of these findings hold. Moreover, Bayesian aggregation confers the advantage of enhancing precision in particular for individual TOT estimates.

We investigate the role of delivery type (volunteer or teacher) within the basic model structure presented above by fitting the Rubin (1981) model to each subset – teachers versus volunteers – separately. We present ITT results split by delivery model in Figure 7 and TOT results split by delivery model in Figure 8. The findings show the importance of the delivery model, especially for the TOT results, where we see even greater pooling due to the even greater similarity in effects. The visual clustering suggests that when we account for implementation (TOT vs ITT effects) and delivery model (teachers vs. volunteers) the treatment effects may be more generalisable. However, this figure alone is only suggestive; we will formalise this investigation in the following section.

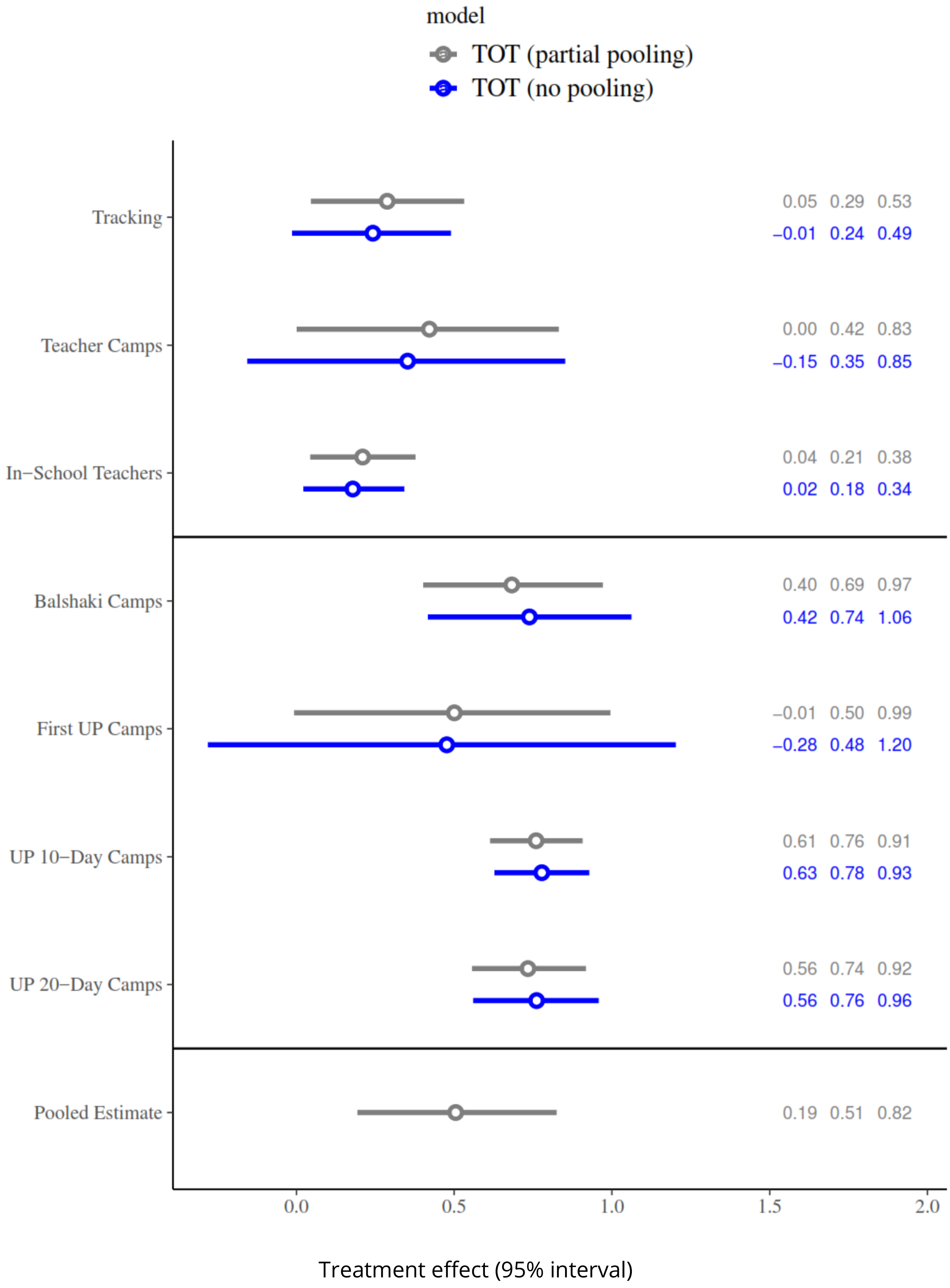


Figure 6: Bayesian aggregation of all TOT results

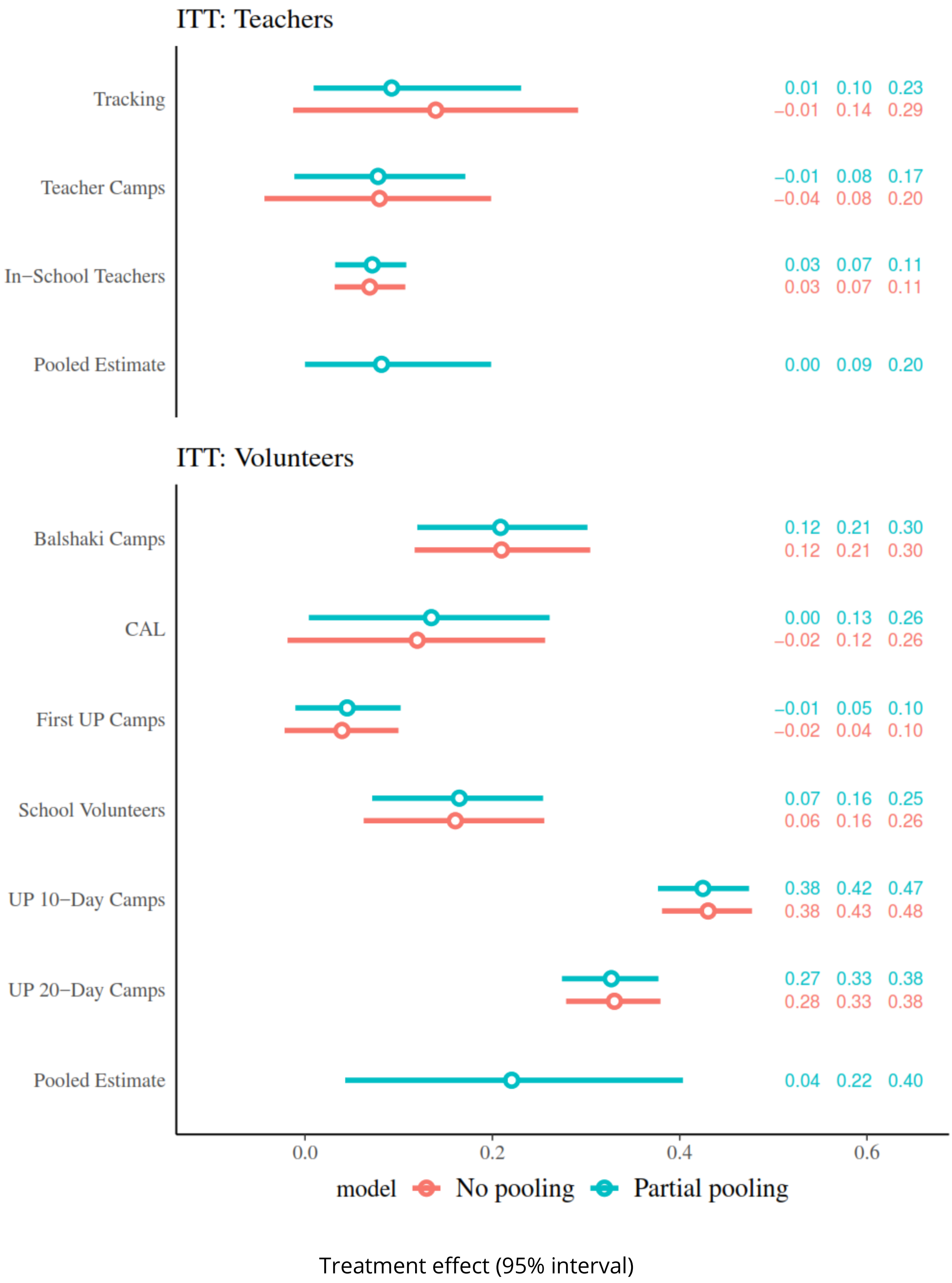
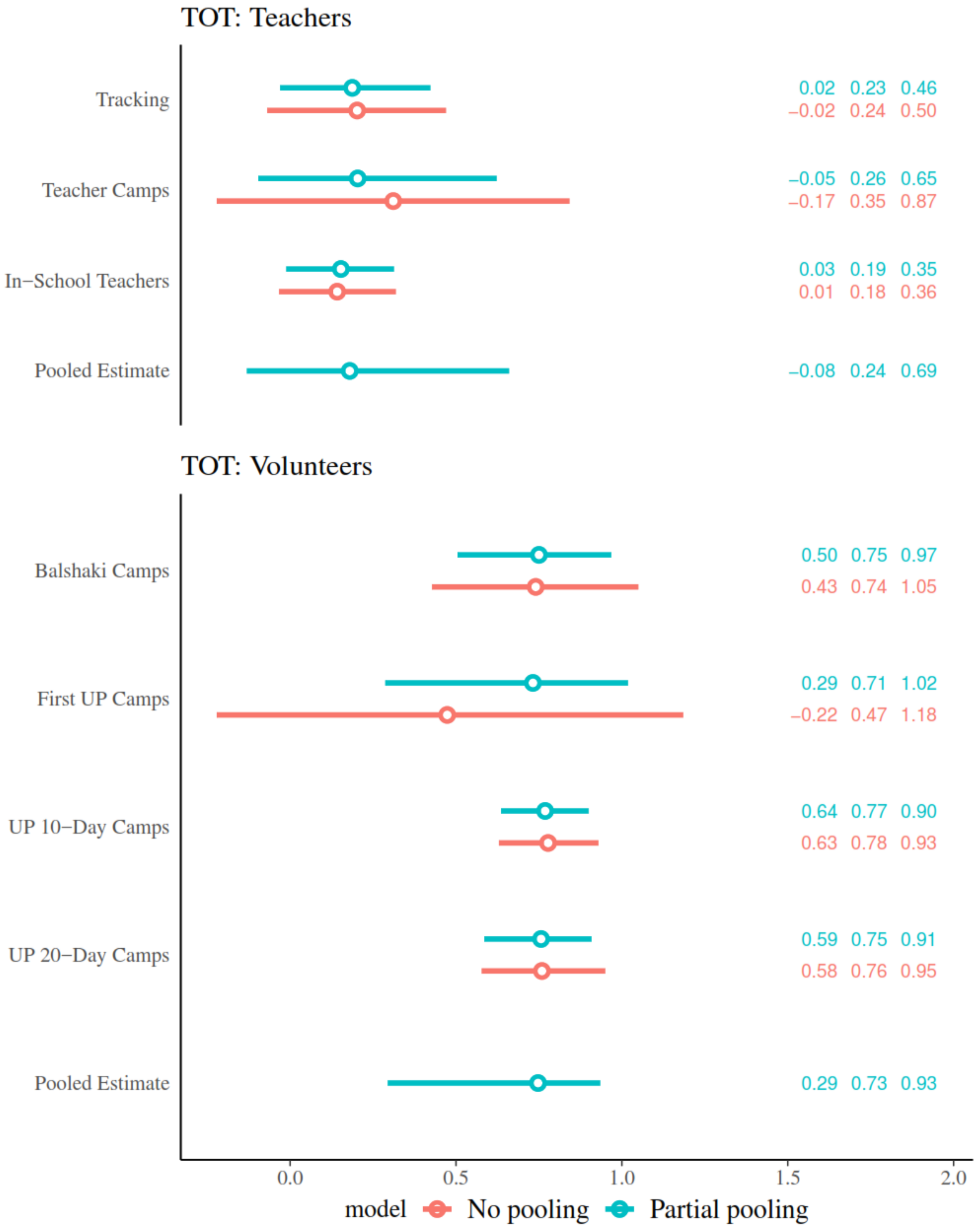


Figure 7: Bayesian aggregation of ITT split by delivery



Treatment effect (95% interval)

Figure 8: Bayesian aggregation of TOT split by delivery

5.2.2 Bayesian Meta-regression Results

To systematically analyse which factors are most closely related to the observed treatment effects of targeted instruction, we now turn to the meta-regression results. We use the model described in Section 4.3, and we consider primarily the baseline education attainment and the delivery model of targeted instruction as our covariates of interest. In Figure 9, we show the results of fitting these models for both the ITT and TOT effects, alongside the original results of the basic aggregation, as well as the results of meta-regression models fit with either only baseline information or only delivery information. We present the inference for each study as well as the pooled estimate (the bottom row of the graphic), which is the average effect of targeted instruction across all settings.

As Figure 9 shows, for the ITT, running meta-regression models conditioning on either or both covariates of interest has little impact on the inference. The basic Rubin (1981) findings are confirmed by the more advanced models: the ITT effects are positive yet substantially heterogeneous across settings. By contrast, the TOT effects are now less heterogeneous, and what heterogeneity is present is substantially explainable by the presence of these covariates. The TOT model simply conditioning on delivery type has both the largest average effect at a little over 0.53 SD, as well as the most precise inference (this is the pink bar on the bottom line of the figure). Examining each study in turn, we can see visual evidence that the Teacher Camps, Balshaki Camps and First UP Camps have their estimated TOT effects somewhat revised upwards. We check robustness and confirm that the results are not contingent on any one of these studies in Appendix B.

The fact that that conditioning on baseline educational performance does little to improve precision is quite surprising. One might expect either a negative correlation due to upper boundary effects (high-performing students or schools already perform well, so benefit less from remedial classes) or a positive correlation due to selection (students in high-performing schools know how to learn, so benefit more from remedial classes). It is possible that in this dataset all students are so far behind the curriculum that the variation in baseline learning is irrelevant to their progress under targeted instruction, but it is also possible that delivery model and implementation (measured via take-up) simply play a larger role in practice.

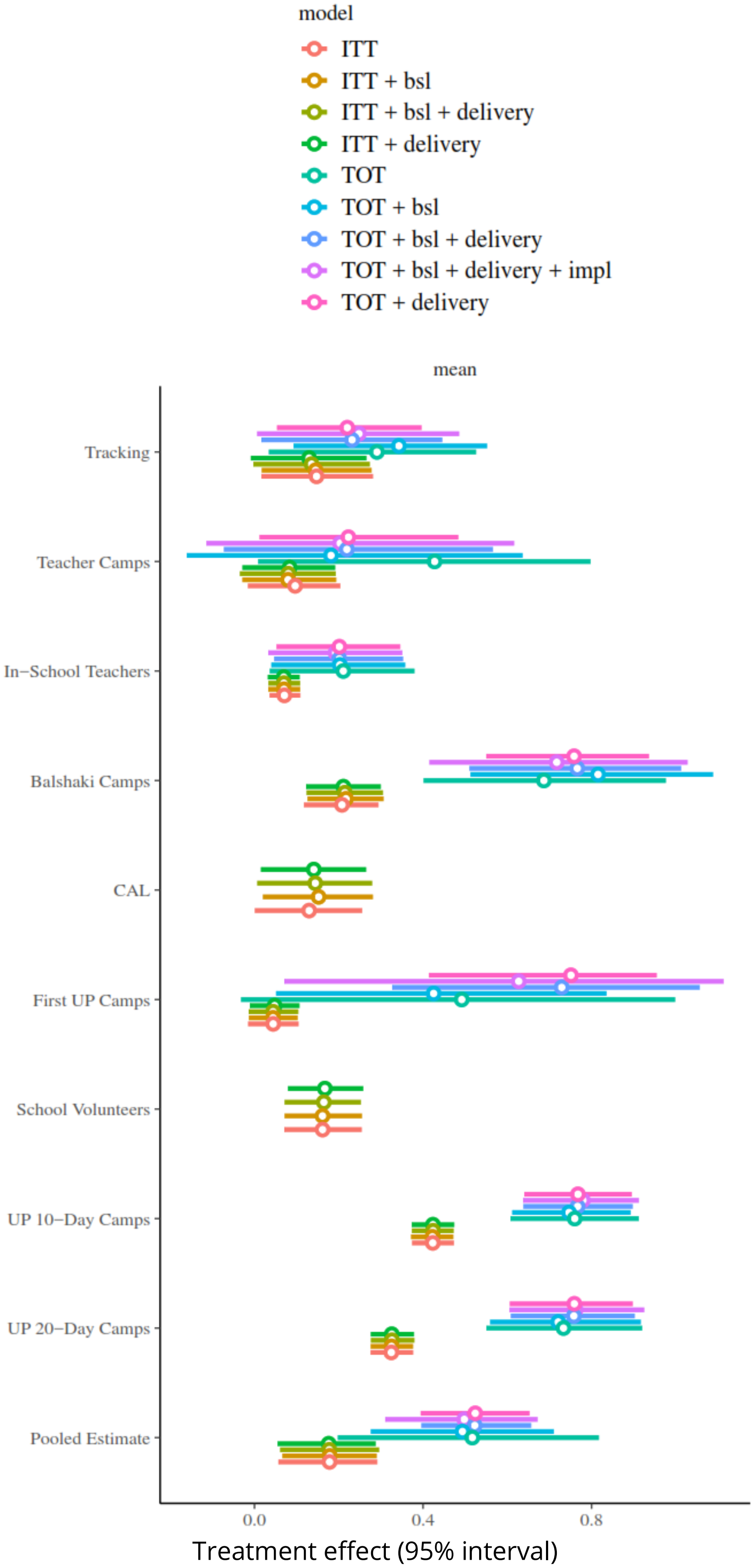


Figure 9: Summary of Bayesian aggregation results across multiple models

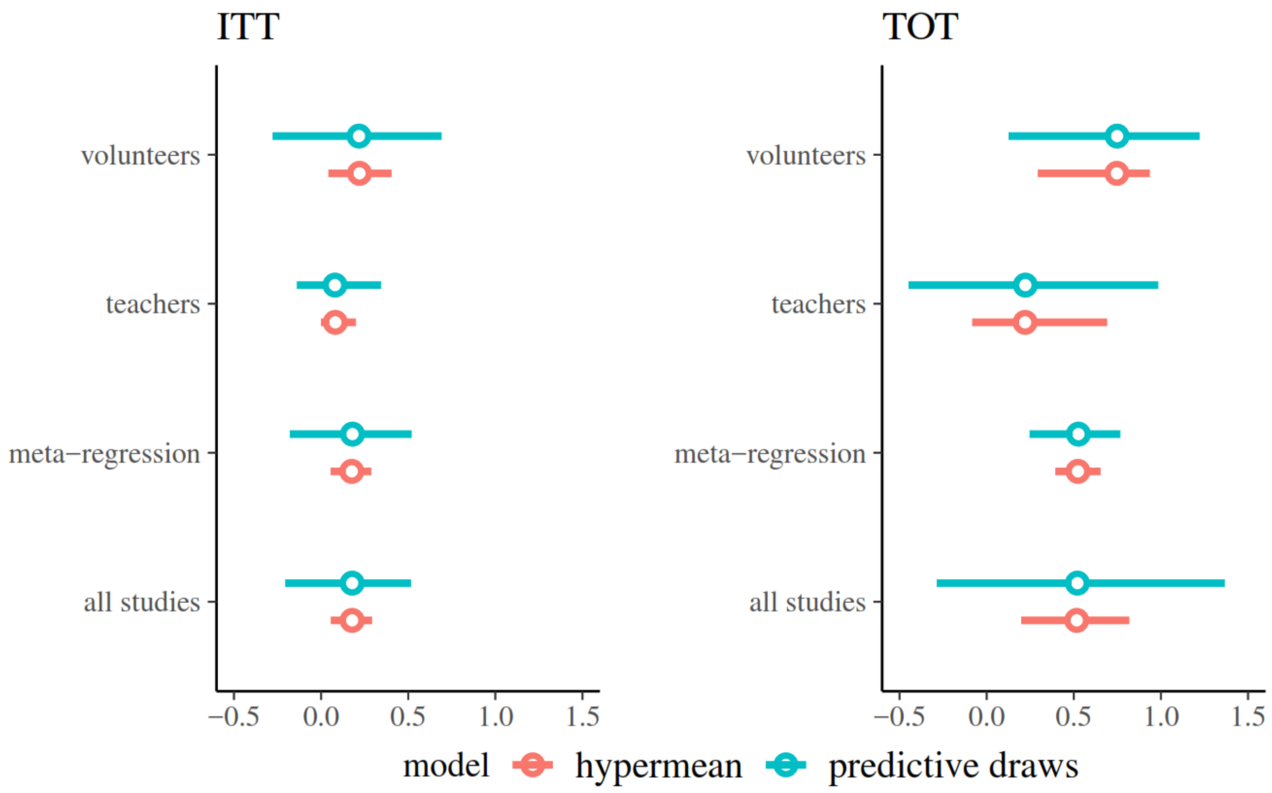


Figure 10: Posterior predictive distributions of future effects

5.2.3 Posterior Predicted Treatment Effects

To understand how our statistical aggregation results translate into extrapolation to future policy settings, we now examine the posterior predicted distributions of the next comparable study's ITT and TOT effects, respectively. Figure 10 shows the uncertainty interval of the predicted effect of targeted instruction in the next setting, labelled 'predicted draws', and for comparison also shows the uncertainty interval on the average effect of targeted instruction across settings.

To interpret Figure 10, recall that if the effect of targeted instruction were homogeneous in all settings, the red and green distributions would be the same because the average effect would then be the predicted effect everywhere. Classical fixed-effects meta-analysis does not distinguish between these two quantities; in that context, the posterior uncertainty on the hypermean is the posterior uncertainty on the predicted effect. But in the presence of heterogeneity of effects across settings, there is a fundamental extrapolation error when we attempt to use the mean to predict the specific effect in any setting, which ought to be reflected in greater prediction uncertainty; this can be captured in the hierarchical model. Our results show that there is heterogeneity in both the raw ITT and TOT effects of targeted instruction, but the gap is much smaller for the meta-regression model on the TOT effects, confirming that accounting for contextual factors eliminates much of the heterogeneity across settings. The figure further shows that accounting for different delivery mechanisms seems to capture some of the variation in effects in the TOT, as the uncertainty is lower on the split models than on the average of all studies, even though the average is estimated from more data.

We observe a few patterns. First, average effects and effects for teachers do not consistently have positive effects in all posterior distributions; only volunteer TOT estimates do. This is likely due to the fact that the volunteer TOT effects are both much larger on average and, crucially, much more homogeneous across studies. In short, our analysis finds strong evidence that volunteer-led targeted instruction interventions have a generalisably large and positive impact. By contrast, though the average TOT and ITT effect for all targeted instruction programmes is positive, in each subcase and in each model there is too much heterogeneity across study contexts to rule out the potential for negative effects in a model that allows for effect distributions to be symmetric (as all classical meta-analytic models do).

Figure 10 also shows that for TOT effects, meta-regression substantially improves the precision of the inference on the hypermean and the posterior predictive draws. As the right panel shows, the posterior predicted TOT effect from the meta-regression model is even smaller than the posterior hypermean of the basic Rubin (1981) model – that is, these covariates more than compensate for the original extrapolation error that one would have attained in the basic Bayesian – or indeed frequentist – aggregation exercise. Moreover, the variation within the delivery groups (teachers and volunteers) is larger than the variation remaining once one conditions on this type of study (meta-regression). These results formally confirm our earlier finding that implementation (TOT vs ITT) and delivery mechanism (volunteer vs teachers) substantially explains variation in TOT results, and has a key role to play in predicting the relative success of targeted instruction interventions across settings.

5.2.4 Results with Theory- and Expertise-Informed Priors

In this section, we combine the findings of the Bayesian statistical analysis presented in the previous sections with qualitative expert insight and economic theory. This approach to understanding generalisability bridges economics and epidemiological practice in a manner consistent with the advice for researchers set out in Deaton and Cartwright (2018) and Vigneri *et al*. (2018). First, we discuss the theory. Second, we present evidence from the literature. Third, we present results using both types of information – theory and expertise based on the literature – captured formally in the model via priors.

Theory-informed priors - We start by formalising components of the theory of change. According to an extensive literature, as well as qualitative expertise, targeted instruction is designed to bridge learning gaps when student learning levels are far behind the expected grade level. In this environment the curriculum is poorly matched to students' zone of proximal development. Moreover, targeted instruction works by creating homogeneous groups, which enables more efficient instruction by minimising the likelihood of mismatch in a given group of receivers of information. This approach is most needed in schooling systems in which, in the *status quo* situation, student learning is far behind the expected grade level and where there is significant heterogeneity, and thus high mismatch between curricula and any given student learning level.

The theoretical framework outlined above predicts that students in lower learning levels are most likely to gain from the targeted instruction intervention. This is consistent with a broader economic notion of diminishing marginal returns. On the other hand, there is related economic theory on the notion of complementarities, whereby adding one activity increases the returns of the other; this is behind much of the 'big push' development literature that underpins many highly influential development programmes, including the Millennium Villages Project and the BRAC Graduation programme (sometimes called 'Targeting the Ultra Poor') (Banerjee *et al*., 2015). This intuition applies to the targeted instruction intervention: for example, once a child can recognise numbers, they can more easily learn to do addition, consistent with the prominent notion of dynamic complementarities in skill formation (Cunha and Heckman, 2007). This suggests that students who start at higher baseline learning levels will progress faster.

Since these theories give rise to qualitative predictions that go in opposite directions, the overall implication for the quantitative estimation model is that one should regularise the correlation between control level (or baseline level) and treatment effects across settings. The translation of the qualitative understanding into a quantitative input via the prior proceeds under the following logic. First, we observe that the two countervailing mechanisms are likely to both be operating in each setting, or at least we do not have any strong reason to believe that one of these mechanisms typically dominates over the other. Next, we observe that if the two mechanisms were of exactly equal strength, the correlation observed in the data between the baseline level of ability and the treatment effect of TaRL would be exactly zero. While there is no basis for believing that the two effects would exactly counterbalance each other, in the absence of evidence that one of these mechanisms overwhelms or dominates over the other, one should not expect to see a large correlation of either sign in the data. This corresponds to a prior that places equal weight on positive and negative correlations but predicts a higher likelihood of moderately sized correlations of either sign than of extreme correlations of either sign; this offers a smooth, classical regularisation in the style of the ridge penalty.

Literature-informed priors - We now use the theory discussed above to inform a set of stronger priors on our Bayesian evidence synthesis. We augment the above discussion with additional information based on the literature on educational interventions in developing countries. To ensure that this information is generated prior to any of the evidence on targeted instruction models contained in our present dataset, we limit ourselves to literature published before 1995 (before Pratham was founded). At that time, the state-of-the-art understanding of experts in the field of primary education interventions in developing countries was overall quite pessimistic about the potential for any single intervention to improve outcomes (Lockheed and Vespoor, 1991). One exception to this general pessimism related to the possibility of providing incentives to teachers, although deeper discussions in the academic literature noted that there seemed to be a potential role for pedagogical improvement, and that the incentives might primarily work to improve pedagogy. However, this potential was largely speculative (Hanushek, 1995; Vespoor, 1989). Overall, the development economics literature was pessimistic about the potential for non-incentive-based reforms to have a major impact on children's learning outcomes.

Thus, the state of the field's understanding prior to targeted instruction models further motivates a reasonably tight prior around a zero effect size. Such a prior encapsulates the qualitative notion that targeted instruction would have to overcome priors against it to prove itself in that intellectual climate. To investigate the results of using strong theory- driven priors, we now present results from the Rubin (1981) models under a variety of much stronger priors than those used in the previous section. Using a range of strong priors allows us to understand exactly how pessimistic one would have had to be before one could disregard the strongly positive empirical evidence from the basic, 'reduced form' statistical analysis that did not incorporate any substantive theory via priors.

Results with theory- and expertise-informed priors - Figure 11 shows the results of the basic aggregation model for the ITT results under a variety of priors on the average effect (hypermean) parameter, and Figure 12 shows the same analysis for the TOT effects. All priors we consider are Gaussian, as discussed in Section 4, and are centred at a zero effect, as discussed above, yet their strength varies substantially by varying the SD of the Gaussian prior around zero. For both ITT and TOT estimates, we show the results of a reasonably strong negative prior on the effect, represented by a prior variance on the hypermean of 0.5 outcome units, an even stronger negative prior represented by a smaller variance of 0.25, and the strongest negative prior with a variance of 0.1. This is extremely tight relative to the default priors we used in the previous section, and which are commonly used in the literature, where the prior variance is more like 5–10 units (e.g. the default used is 7.8), not 0.5–0.1, as informed by the theory discussed in this section.

Figure 11 and Figure 12 show that the evidence on the positive ITT of targeted instruction is extremely strong across all priors, while the TOT results are somewhat more influenced by the priors. This is because the TOT effects are estimated with greater uncertainty within each study, and therefore are less able to overcome pessimistic priors. Yet all but the most pessimistic priors report an almost certain positive TOT effect on average, and even with the most pessimistic priors, the Bayesian models report a more than 75% chance of a positive TOT effect of targeted instruction. In Appendix A3, we also show results using a model that implements regularisation of the correlation between baseline levels and treatment effects. We find similar patterns: the strength of the TOT evidence on a positive average (hypermean) is shown in the compensating pattern in the hyperSD; if the hypermean is forced down closer to zero, the hyperSD is forced upwards to compensate for the fact that we have strong evidence of some effects being large and positive in some studies.

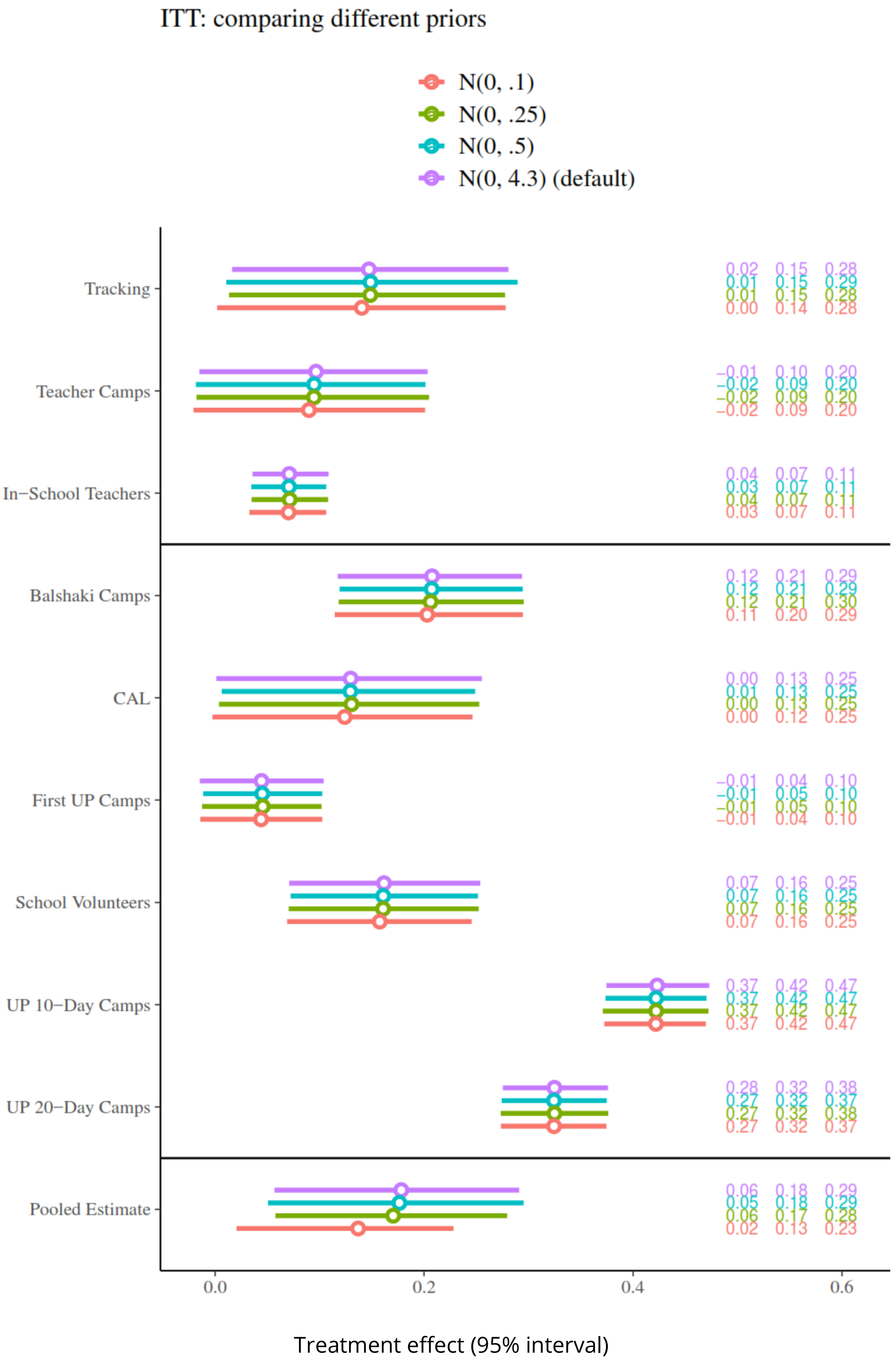


Figure 11: ITT results under different theory-driven priors

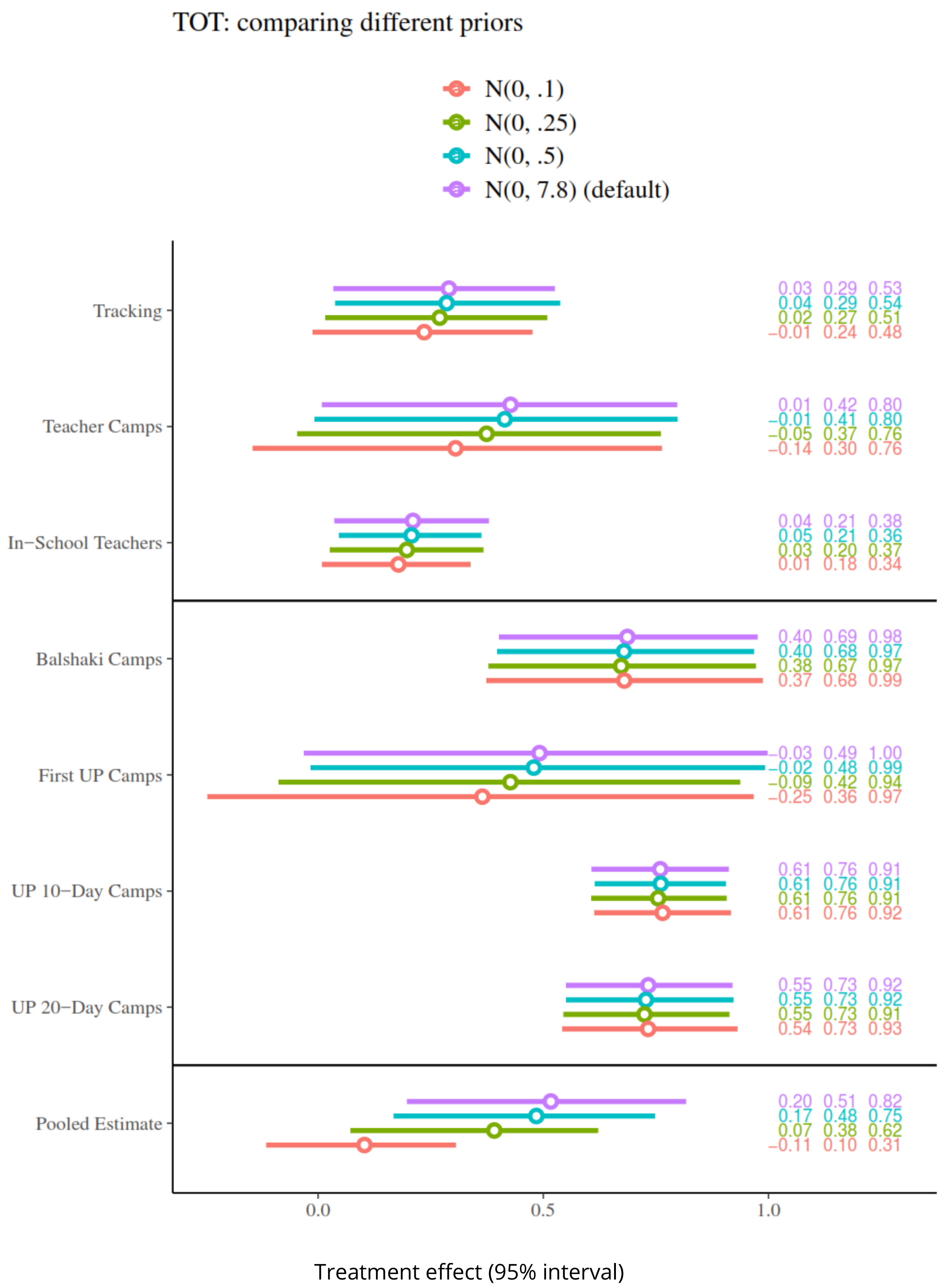


Figure 12: TOT results under different theory-driven priors

Overall, the evidence on the positive impact of targeted instruction is strong even when we impose strong priors, suggesting the patterns in the data are robust and informative. This result aligns with the progression of expert opinion. While in the 1990s views were pessimistic regarding the potential effectiveness of non-incentive-based education reforms, after decades of rigorous evidence, an emerging view is that pedagogy reforms, rather than resource or incentive reforms, are the most promising (Global Education Evidence Advisory Panel, 2020). This shift in opinion had to overcome strong priors, and our analysis shows the evidence generated by targeted instruction is indeed strong enough to achieve this.

6. A Model: The Role of Implementation in Identification of Treatment Effects and Generalisability

Our evidence synthesis reveals the central role of implementation in generalising the effects of targeted instruction. Motivated by these results, we develop a new Bayesian hierarchical likelihood model that captures the importance of programme implementation. This model accounts for uncertainty around implementation, as well as sampling variation and treatment effect variation.

We first define our notion of implementation, which we conceive of as a factor that scales the latent treatment effect to generate the actual, realised treatment effect in a given setting. We then show that in the absence of information about programme implementation, neither the latent programme impact nor its generalisability across settings can be formally not identified. We show that introducing data on programme implementation allows for both identification and estimation of the average impact and variance across settings in a hierarchical structure. We then test the model using simulated datasets to objectively measure its performance. Finally, we fit the model to targeted instruction data, capturing two implementation dimensions: take-up and fidelity. The results show that the latent treatment effect of targeted instruction is expected to be more than double the realised treatment effect reported in the current literature.

6.1 Defining Implementation

Consider a set of contexts $j = 1, 2, 3 ... J$. In each setting, there is a latent treatment effect of the programme that is achievable if it is fully implemented, denoted $\theta_j \in$ ı. We do not intend this to capture perfect implementation in every detail; rather, we conceive of 'full' implementation as referring to when the theoretically core components of the programme are indeed delivered to the intended programme recipients on time.

We define a notion of implementation that is a proportion rather than strictly binary, since social programmes often have multiple core components and may reach some fraction of their programme goals even if they do not meet all of them. This is analogous to the way in which programmes may reach some fraction of their intended recipients, a notion already embedded in the economics literature and in our analysis thus far, captured by programme ‘take-up’. This variable is the proportion of recipients who actually receive the programme and defines the wedge between ITT effects and TOT effects.

We consider two core components of implementation: fidelity to programme quality, as well as pure take-up (i.e. attendance). In the case of targeted instruction, fidelity means that teachers assess the children’s learning level, group the children by their level, and then instruct them at their level. When these three things did not occur – and as both our data and our field experience shows, these three things are not trivial to execute – then ‘targeted instruction’ did not happen. Hence, we need a notion of implementation that is broader than attendance or take-up, and that also captures the proportion of instances in which the core components of a programme were executed effectively.

We thus define the degree to which a programme is implemented in context j as a proportion $m_j \in [0, 1]$, which we call the ‘implementation factor’. This factor m_j is 0 when no component of the programme is delivered to recipients, and m = 1 when the programme is delivered as intended in every aspect to all recipients. When implementation is only partially achieved, one might expect to only receive some corresponding fraction of the impact on recipients that one would have received if the programme had been perfectly implemented. For example, consider defining implementation level as the percentage of instructors who grouped students by ability. This percentage is the implementation level m_j, and the latent θ_j would be the latent treatment effect of receiving instruction which is actually targeted to the right level.

To summarise this formally, we consider a set of programme contexts indexed by $j = 1, 2, 3...J$ and offer three definitions of relevant objects.

Definition 1: Implementation factor *The implementation factor, denoted $m_j \in [0, 1]$ for a setting j, is the extent or proportion to which the programme was effectively implemented in setting j.*

Definition 2: Latent treatment effect *The latent treatment effect (LTE$_j$), denoted $\theta_j \in \mathbb{R}$ for a setting j, is the impact achievable when the programme is fully implemented.*

Definition 3: Realised treatment effect The realised treatment effect (RTE$_j$) is the observed impact of the programme in setting j, defined as:

$$RTE_j \equiv m_j \theta_j$$

Randomised trials recover observed treatment effects, which we define as the RTE_j. We define this treatment effect as the *LTE* θ_j multiplied by the implementation factor m_j. This structure is analogous to the definition of the ITT effect, which is the TOT scaled by take-up proportion.[13]

This is a multiplicative implementation model: if the implementation factor m_j is less than 1, one should not expect to obtain the same effect as if the programme had been fully implemented. Instead, one should expect to have an impact that is only a fraction of the latent potential effect: $m_j\theta_j < \theta_j$. Our model intentionally allows the implementation factors to vary across studies, since these factors are just as likely to be influenced by contextual factors as the underlying LTEs are.

6.2 Identification

In this section we establish that the presence of the implementation factor m_j means that LTE in any given setting cannot be identified from realised effects if implementation is unknown. To make this point clearer, For intuition, consider the most obvious case when this causes a problem: null effects. Null effects could be due to an ineffective programme or an effective programme which was never implemented. Without information on implementation, we could misattribute a null effect to a treatment effect when in fact it is null implementation. More formally, a programme that has no treatment effect ($\text{RTE}_j = 0$) could be driven by a situation in which $\theta_j = 0$, but, equally possibly, $m_j = 0$. Without explicit information on m_j, a realised effect of zero cannot logically be used to infer a null LTE θ_j. In other words, the underlying effect θ_j is not identified from the data. While this lack of identification is to some extent intuitively evident, we find it useful to lay out the formal result to confirm our intuition and to emphasise that the relevant parameters can be estimated.

Following Lewbel (2019), we define (point) identification of any parameter as the values of a parameter producing observable distributions of data. We show that the latent effect of a programme, θ_j , is not identified from the programme treatment effect RTE_j when implementation is not known.

[13] A series of similar assumptions apply in this model as apply in TOT estimation to extrapolate effects to the broader population, such as that the sample receiving the programme should be statistically equivalent to the broader sample on covariates (i.e. not selected). In the targeted instruction case, this appears likely, with implementation occurring on average with the majority of the sample.

Definition 4: Identification *A parameter θ is point identified from some observable statistic $\phi(\theta)$ or distribution of data $F(\theta)$ if for any $\theta' \neq \theta$, $\phi(\theta') \neq \phi(\theta)$ and $F(\theta') \neq F(\theta)$.*

Proposition 1 *If implementation m_j is not observed, the latent treatment effect of any programme, θ_j, is not identified even if the realised treatment effect RTE_j is observed.*

Proof From Definition 3, the RTE identified in a randomised trial is $TE_j = m_j\theta_j$. In the absence of information about m_j it is possible that $TE'_j = TE_j$ even when $\theta'_j \neq \theta_j$. Suppose that $\theta'_j = a\theta_j$. If $m'_j = \frac{1}{a}\, m_j$ then $\theta'_j\, m'_j = a\theta_j * \frac{1}{a}\, m_j = \theta_j\, m_j$. From Definition 4, this means θ_j is not identified. ■

While the identification result above is general, the most concerning possibility it presents is that of false negatives in treatment effect attribution. If we do not have data on programme implementation, an observed null effect could be misattributed to an intervention not being effective, when in fact it was never actually implemented. Even if we have data on programme implementation, if $m_j = 0$ then $RTE_j = 0$ for any θ_j. In this case the LTE of the programme is not identified even when implementation is observed. In summary, lack of implementation information or extremely poor implementation makes it impossible to attribute effects to the treatment, and is a major threat to the internal validity of a study.

We now show that a similar lack of identification affects the generalisability of the evidence on programme effects across settings when implementation is not recorded across settings. Recall that we have study settings $j = 1, 2, 3 \ldots J$, each with their own tuple (m_j, θ_j, TE_j), and variance is therefore defined across settings.

Proposition 2 *When $\{ m_j \}_{j=1}^{J}$ is not recorded, the heterogeneity in the set of RTE $\{ RTE_j \}_{j=1}^{J}$ does not identify the heterogeneity in the set of LTE $\{ \theta_j \}_{j=1}^{J}$ even when implementation is homogeneous.*

Proof From definition 3, $RTE_j = m_j\theta_j$, so

$$
\begin{aligned}
var(\, RTE_j \,) &= var(\, m_j\theta_j \,) \\
&= E[(\, m_j\theta_j \,) - E[\, m_j\theta_j \,]]^2 \\
&= E[\, m_j^2\theta_j^2] - E[\, m_i\theta_i \,]^2 \\
&= Cov(\, m_j^2, \theta_j^2) + E[\, m_j^2]E[\theta_j^2] - E[(\, m_j\theta_j \,)^2]
\end{aligned}
$$

Since all m_j are not known, this does not identify $var(\, \theta_j)$. To see this more easily, consider $m_j = m$ $\forall\ j$, the case of perfectly homogeneous but still unknown implementation across settings. Then,

$$var(\ RTE_j) = m^2 var(\ \theta_j)$$

But since m is not observed, it is possible to have the same $var(\ RTE_j)$ reflect different $var(\ \theta j)$. Specifically if $\theta'_j = \sqrt{a}\theta_j\ \forall\ j$, and $m' = \frac{1}{\sqrt{a}} m$, then:

$$\begin{aligned} var(\ \theta'_j m'\ ) &= (\tfrac{1}{\sqrt{a}} m)^2\ var(\sqrt{a}\theta_j) \\ &= \tfrac{1}{a}\ (m)^2 * a * var(\theta_j) \\ &= m^2 var(\theta_j) \\ &= var(\theta_j m).\ \blacksquare \end{aligned}$$

The problem arises because anything, even a fixed number with no variation in itself, which scales a random variable's magnitude also scales its variance. Intuitively, the problem is worse when implementation is heterogeneous. Even if the variation in implementation is independent of variation in potential effects, the presence of this extra variation makes the programme look less generalisable than it is. If variation in implementation is positively correlated with potential effects, the distortion is even greater; if the correlation is negative, the distortion can be reversed. Thus, failure to report implementation means that the heterogeneity in the potential effects is not identified.

Since we always observe the RTEs, it may be tempting to wonder how serious this identification problem on LTEs really is. Perhaps it is only LTEs, and not latent effects, that really matter in practice, or for policy decisions. Certainly for those who received the programme in the past, the realised effect is all that matters. But for potential future recipients in other contexts, where implementation may be different, the realised effect in previous studies may not be relevant at all. Even for future recipients in the very same context, the RTE only captures all relevant information if we assume the implementation cannot be influenced or changed. But this is not true: implementation level is itself a random variable that researchers and policymakers can affect. We see substantial variation in implementation factors in practice in our dataset, and in Section 7 we will show that it is possible to change the implementation level in the field, with substantial impacts on realised effects. Thus, it is important to jointly study implementation level, LTEs – which capture the full potential for impact – and RTEs together. To achieve this, we require a model that can disentangle these multiple sources of variation. It is this to which we now turn.

6.3 Bayesian Model and Estimation Performance

We now embed the notion of programme implementation developed above into the Bayesian hierarchical aggregation framework, so that the implementation factors $\{m_j\}_{j=1}^{J}$ can formally enter the analysis. This is desirable for two reasons: first, the level of implementation can be correlated with the potential treatment effect and joint analysis of potentially correlated random variables is always preferable, and second, implementation levels often lie near the boundary of the parameter space and extra care is required to infer them. Fortunately, we show in simulations that as long as implementation is not exactly zero and we do have some information about the degree of implementation – even if it is uncertain – then it is possible to still identify the *LTE* θ_j , as well as the variation in this effect across settings, even when *J* is small.

We build our hierarchical implementation factor model from an adapted Rubin (1981) model, incorporating our model of the RTE as the product of the latent potential effect θ_j and the associated implementation factor m_j combined. Because the implementation factor and latent effects are multiplied together, and we observe neither of them directly, we need to perform a statistical deconvolution to identify their distributions separately. We observe the estimated realised effect $\widehat{RTE}_j$ with some noise $\hat{se}_j$ We also observe an estimate of the implementation level, $\hat{m}_j$ with standard error $\hat{se}_{mj}$. We now have to infer the true m_j and θ_j from the data jointly. We can do this using the model below:

$$\begin{aligned} \widehat{RTE}_j &\sim N(m_j\theta_j, \hat{se}_j^2) \\ \hat{m} &\sim N(m_j, \hat{se}_{mj}^2) \qquad \text{(Equation 6.1)} \\ \theta_j &\sim N(\theta, \sigma_\theta^2) \end{aligned}$$

To make estimation of this model concrete, consider the definition and measurement of the implementation factor for targeted educational instruction programmes. A researcher could define the implementation level purely as student take-up; in this case the recorded attendance rate of the classes would form the estimate $\hat{m}_j$, and the LTE θ_j would be the TOT effect we are familiar with. Or, if we instead define the implementation level as the percentage of instructors who grouped students by ability, then this percentage would form the estimate $\hat{m}_j$, and the latent θ_j would be the effect of receiving instruction which is actually targeted to the right level.

A natural next question, however, is how to define implementation when we have data on multiple aspects of programme execution. The natural answer is to apply the 'm-factor' logic recursively: let us say attendance of students in programme *j* is captured by a variable $m1_j \in [0, 1]$ and fidelity of instruction is captured by another variable $m2_j \in [0, 1]$. Logically, if only half the students show up, this dilutes the effect that the programme can have by half – and if only half the instructors actually deliver targeted instruction, this dilutes the programme effect by half again. To perform joint inference on all these factors, the following model may be used:

$$\begin{gathered} \hat{RTE}_j \sim N(m1_j m2_j \theta_j , \hat{se}_j^2) \\ \hat{m}1_j \sim N(m1_j , \hat{se}^2_{m1_j}) \\ \hat{m}2_j \sim N(m2_j , \hat{se}^2_{m2_j}) \\ \theta_j \sim N(\theta , \sigma_\theta^2). \end{gathered} \qquad \text{(Equation 6.2)}$$

This 'two-factor' form of the model allows us to make progress not just on understanding whether or how much implementation matters but which aspects of implementation matter. Conceptually, the model may be expanded to as many factors of implementation as the analyst thinks necessary, as long as one can get data on them. A drawback of the two models offered above is that they do not explicitly consider correlations between implementation levels m_j and latent effects θ_j – this amounts to assuming that m_j carries no additional information about θ_j after they have been deconvolved, such that places with higher m_j are not systematically different in terms of their θ_j. This assumption simplifies the models enough to make them tractable even on small datasets. Whether this simplifying assumption holds is an empirical question, and the likelihood of this assumption holding could be addressed by considering a richer model with a joint hierarchical structure placed on (m_j , θ_j).

We now show via simulation that reliable estimation and inference is possible using both the one-factor model and the two-dimensional m-factor model even when J is quite small. We consider datasets of size $J = \{3, 5, 8, 15\}$, and for each case we run 250 simulations from the model above, where the true hypermean is 10 and the true hyperSD is 7. We draw the J standard errors on the RTEs from a uniform distribution from 10 to 20. We draw the J true implementation factors from a uniform distribution on [0.1, 0.9], which is the range in our dataset, and we draw their standard errors from a uniform distribution on [0.005, 0.05] because this is roughly their magnitude in our dataset. In each case we record the root mean squared error (RMSE) of the posterior mean and posterior median of each of the hyperparameters (θ , σ_θ), as well as the true frequentist coverage of the 50% and 95% posterior credible intervals across the 250 simulations for each case.

The results for the single implementation factor model (Equation 6.1) are shown below in Table 2, and in Table 3 for the two-factor implementation model (Equation 6.2). As the results show, the 95% Bayesian credible interval typically has greater-than-nominal frequentist coverage at all values of J. However, in the two-factor model, the 50% credible interval's coverage is degraded for the HyperSD when $J < 15$. The results show that using the posterior median offers large RMSE gains for the hyperSD relative to the posterior mean, and roughly comparable RMSE for the hypermean. The improved performance of the posterior median is likely due to the inherent skewness of the posterior distribution of the hyperSD. Overall, the reasonably low RMSE for $J > 3$ offers assurance that the greater-than-nominal coverage of the credible intervals is not due to these intervals being unduly wide, though we certainly see gains from collecting more studies.

Table 2: One-factor implementation model performance in simulations

Studies	Parameter	RMSE (mean)	RMSE (median)	50% CI coverage	95% CI coverage
J= 3	Hypermean	7	5	0.996	1
	HyperSD	66	30	0.016	1
J= 5	Hypermean	3	3	1	1
	HyperSD	16	8	0.632	1
J = 8	Hypermean	2	2	0.980	1
	HyperSD	6	3	0.984	1
J = 15	Hypermean	2	2	0.968	1
	HyperSD	1	1	0.996	1

The two-factor implementation model is conceptually preferable, but the single-factor model performs better when *J* is small. Hence, in our results, we rely on the single-factor model. Even if we conceive of implementation as take-up, using the single implementation factor model is preferable to first computing the TOT using an IV strategy or Wald estimator and then aggregating the result. This is primarily because joint analysis allows us to deconvolve the whole distribution, not just the expected value of the treatment, and thus we can account for uncertainty in implementation, rather than conditioning on it via an inputted standard error on a TOT estimate. In addition, we prefer this approach since take-up in our data ranges from 0.08 to 0.90, and the Local Average Treatment Effect (LATE) is underpinned by a linear probability model, which is unlikely to perform well over extreme values.

Table 3: Two-factor implementation model performance in simulations

Studies	Parameter	RMSE (mean)	RMSE (median)	50% CI coverage	95% CI coverage
J = 3	Hypermean	15	5	1	1
	HyperSD	167	85	0	0.936
J = 5	Hypermean	4	3	1	1
	HyperSD	44	27	0.016	1
J = 8	Hypermean	3	3	1	1
	HyperSD	19	12	0.196	1
J = 15	Hypermean	2	2	1	1
	HyperSD	7	4	0.944	1

6.4 Implementation Model Results

We first fit the single-factor implementation model (Equation 6.1) to our data considering take-up of the programme as the level of implementation, as this variable is observed in seven studies. Table 4 shows the results for all studies in Panel A, teacher-delivery method studies in Panel B, and volunteer-delivery method in Panel C. We show the posterior mean along with five posterior quantiles to give the full sense of the distribution, and report the Rhat criterion as a convergence diagnostic. As the results show, the LTEs for all studies are both much larger than the average realised effects and more generalisable, but the difference is much more marked for volunteer studies. The average LTE for volunteer-delivered programmes is 0.47 SDs, compared to 0.22 SDs for teacher-delivered programmes. Per our simulations, we use the posterior median of the hyperSD as our preferred estimator for this parameter, and we find approximate hyperSD of 0.15 SD units for each of the delivery models; this implies that LTEs for teachers are likely to be positive in most settings, whereas for volunteers they are always large and positive.

Table 4: Implementation model posterior intervals: inference on take-up

	mean	2.5%	25%	50%	75%	97.5%	Rhat
Panel A: LTEs (all)							
Hypermean	0.394	0.205	0.336	0.390	0.447	0.605	1.003
HyperSD	0.203	0.066	0.129	0.181	0.251	0.483	1.001
Panel B: LTEs (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: LTEs (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 Markov chain Monte Carlo (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 Section 6.3.

We note that the inferred LTEs are somewhat smaller than the direct TOT analysis results from previous sections. This is likely to be because this model accounts for uncertainty on implementation during the aggregation process, rather than conditioning on it as part of a reported standard error on the estimate. This introduces more uncertainty and allows the priors to regularise the estimation towards zero to a somewhat greater extent, as is appropriate in small samples. We also note that the level of take-up in our data ranged from

8% to 90%. Given this substantial range, the linear probability model that underpins the Wald estimation of the TOT effects is likely to be somewhat stressed by the data, and perhaps unduly influenced by extreme results. Our model places bounds on the implementation factors' values without imposing a linear probability model, which may be another reason why we see more uncertainty in these results.

We now consider the data on programme fidelity as another important aspect of the implementation of targeted instruction. Although we only have these data for three study arms and we view the results below as suggestive, the single-factor implementation model still performed well in simulations at J = 3 and the coverage of the 95% posterior interval on the two-factor model was decent (see Table 3). We consider it appropriate to proceed with caution. Table 5 shows the results of fitting the single-factor implementation model (Equation 6.1) to the TOT estimates with fidelity as the implementation level in Panel A, and Panel B shows the results of fitting the two-factor implementation model (Equation 6.2) to take-up and fidelity jointly. In both cases, we see that the potential for additional large upsides to the LTEs is considerable: the results in Panel A show the median LTE is now around 0.81 SDs, with a hyperSD around 0.21. The inference in Panel B is somewhat more aspirational, but the findings align with Panel A in showing even larger LTEs once fidelity is accounted for. Comparing the results in Panel A of Table 5 to the results of considering only take-up (Panel A of Table 4), we find an additional 0.4 SD improvement in the LTE. This is double what we find when we only consider take-up, suggesting that implementation of targeted instruction is indeed more complex than simply ensuring recipients turn up for the classes.

Table 5: Implementation model posterior intervals: inference on fidelity and take-up

	mean	2.5%	25%	50%	75%	97.5%	Rhat
Panel A: Fidelity on TOT							
Hypermean	0.807	0.199	0.718	0.826	0.916	1.297	1.014
HyperSD	0.348	0.009	0.091	0.212	0.445	1.392	1.011
Panel B: Fidelity and take-up jointly							
Hypermean	1.199	0.032	0.989	1.137	1.357	2.645	1.004
HyperSD	0.804	0.012	0.134	0.390	1.032	3.847	1.005

Note: This inference is suggestive as it is generated by three 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 Section 6.3. The joint model results in Panel B should be treated as suggestive because model performance is not reliable for J = 3, although the 95% interval coverage is above nominal.

7. Implementation Can Be Improved: Evidence from A/B Testing in Botswana

The results of our evidence aggregation establish the importance of implementation in determining programme results and generalisability across settings. This offers suggestive evidence that if implementation can be changed in practice, the gains in children's learning may be substantial. We investigate whether there are concrete ways to increase the take-up and fidelity of targeted instruction in the field. We test approaches to increasing the fidelity of targeted instruction in the context of TaRL in Botswana, where the government is actively scaling up and testing the programme, in partnership with Youth Impact, one of the largest NGOs in the country.

7.1 Intervention and Study Design

In Botswana, TaRL is implemented primarily by grouping students by operation level: that is, whether they can add, subtract, multiply, or divide, or do no operations at all (referred to as 'beginner'). At baseline in our sample there is a lot of variation and low performance along this dimension. Table 6 below shows the highest operation a child can do at baseline in Term 1 of the school year in 2020. Of Grade 3–5 students, 30.4%, can do no operations ('beginner' level), 28.4% can do up to addition, 19.9% can do up to subtraction, 15.4% can do up to multiplication and only 5.9% can do up to division. As Table 6 shows, however, in this sample of students there is variation along other relevant proficiencies as well, such as the ability to recognise and interpret larger-digit numbers. While 22.7% of students can recognise up to four digits, most children cannot, with 45% recognising only up to three digits, and 29.3% of students able to recognise up to two digits.

The lever that was selected to increase the fidelity of the intervention was to increase the likelihood that children will receive instruction that is optimally targeted to their learning level. To test the viability and benefits of such optimisations of targeting instruction to a child's learning level, Youth Impact conducted an RCT comparing two options to subgroup students, a procedure that Youth Impact internally refers to as an 'A/B test'. The standard implementation of TaRL in schools in Botswana ('Option A' in this trial) involves testing and grouping students according to their understanding of operations and then running operation-specific classrooms (e.g. an addition class in one room and multiplication in the next). This means that the operation-level classes occur with student groups who have mixed number recognition abilities. For example, addition-level students who recognise three digits would be in the same small group as addition-level one-digit students. The new treatment being trialled randomly in this sample ('Option B') involves additionally subgrouping students within an operations-level classroom according to their digit-recognition level. For example, addition-level students who recognise three digits would be separated from addition-level students who recognise only one digit, and the instruction is targeted to their digit-recognition level.

Table 6: Botswana sample: learning levels at baseline

Operations	Proportion of students
Beginner	0.30
Addition	0.28
Subtraction	0.20
Multiplication	0.15
Division	0.06
Number recognition	
Zero digits	0.00
One digit	0.03
Two digits	0.29
Three digits	0.45
Four digits	0.23

The trial took place with over 1,000 students across 52 classes in four regions in Botswana, randomised at the class level. While the results of our evidence aggregation offer some hope that this improved targeting may improve learning even more than standard implementation, it is not obvious ex-ante. First, the relationship observed between implementation and effect size across studies could be driven by omitted variables rather than causal from implementation to effect (perhaps certain environments are both easy to implement in and also very suitable for targeted instruction). Second, standard implementation was reasonably high in Botswana and in this context it is not obvious ex-ante that similar-level subgroups will improve learning outcomes beyond the standard classroom-level operation groupings. It is entirely possible that there are diminishing returns to targeting instruction – once instruction is sufficiently targeted, perhaps there is no need to target instruction further, and we may not see additional gains from improved fidelity in this context.

7.2 Results

Table 7 reports the results of the trial, with the data analysed using a standard linear regression model estimated via ordinary least squares. The results show that additional sub-learning-level grouping improves number recognition by 0.21 SDs on average (column 1), with enhanced precision and an effect of 0.22 SDs (p-value <0.05) when controlling for multiple characteristics, such as region and baseline learning levels (columns 2 and 3 show different controls). These effects are considered large in the education literature where successful programmes have effect sizes of typically around 0.10 SDs. Moreover, this effect size is nearly the same size as the gap between ITT and TOT effects in the literature, as well as the difference between the basic Bayesian aggregation model and implementation model, revealing consistent estimates in this randomised trial with those observed in the meta-analysis evidence aggregation. These results reinforce the value of increasing implementation take-up and fidelity, and underline that implementation is not a black-box: rather, improving implementation can be rigorously studied, concrete, tractable, and high-return.

Table 7: Results of a randomised increase in programme fidelity

	Outcome: Number Recognition		
	(1)	**(2)**	**(3)**
Treatment: Sub-level grouping	0.205 (0.160) [0.205]	0.225 (0.099) [0.027]	0.223 (0.097) [0.026]
Baseline number recognition		0.611 (0.053) [0.000]	0.616 (0.054) [0.000]
Observations	1069	1069	1069
Baseline-level controls	No	Yes	Yes
Region fixed effects		No	

Note: All standard errors are robust and clustered at the class level. P-values are reported in brackets. Learning gains are expressed in terms of SDs using the control group SD.

The marginal cost of the targeted instruction optimisation in this trial is small, estimated at just a few cents. As a result, in regard to the cost-effectiveness of optimising targeted instruction, it ranks among the most cost-effective educational interventions, based on a review of over 150 impact evaluations in education (Angrist *et al.*, 2020). Enhancing implementation fidelity may be a particularly efficient use of resources for governments, and for educational approaches that are designed for delivery at scale.

8. Conclusion

The results of our analysis demonstrate the importance of quantifying programme implementation with as much care as we typically apply when quantifying programme effects. We find that implementation levels and delivery modes explain most of the variation in effects of targeted instruction across settings and that this leads to actionable insights that can improve the effectiveness of a programme that is being scaled up. Further study of programme implementation would seem to be promising. This would necessitate the collection of data on take-up and fidelity at a much more detailed and extensive level than currently takes place. This paper reveals that research on implementation is tractable, offers meaningful insights about not just average effects but also, crucially, the generalisability of effects, and can reveal concrete mechanisms that can be used to achieve the largest frontier effects identified in the literature.

Our results suggest several avenues for future work on the effectiveness of targeted education instruction. One question for practitioners in this area is why volunteer-led programmes appear to be so effective relative to teacher-led programmes, even – and especially – when accounting for implementation. The role of teachers and school administrations in the success of these interventions is likely to be complex and deserves greater attention. A related open question is why targeted instruction approaches have been effective in low- and middle-income contexts yet have more mixed results in high-income settings. This could be relevant to the literature on 'differentiated instruction' (Tomlinson, 2014), in which tracking approaches in high-income settings have often been found to be less consistently effective. Programme implementation seems likely to play a role: a recent systematic review highlighted that in many high-income settings 'differentiated instruction has been operationalised in many different ways' (Smale-Jacobse *et al*., 2019). Quantifying the degree of fidelity in targeting approaches in future experimental studies might shed light on this and bridge the gap between results across settings.

Finally, we note that carefully defining and measuring the level of programme implementation may be central to evidence aggregation more broadly. While there are always many potentially important contextual covariates than one can incorporate into meta-analysis, the extent to which the study participants actually received the core components of the programme being studied seems important, yet is rarely measured or accounted for in practice. Our results reveal that doing this can have substantial implications as regards understanding programme effectiveness and generalisability. Moreover, our results offer confidence that accounting for programme implementation can be as simple as aggregating evidence on TOT effects, which is far preferable to ignoring this information. It is possible to go further and to aggregate results using a model which jointly considers evidence on both the implementation levels and LTEs. Finally, implementation can and should be conceived of broadly, including both take-up and fidelity. Developing new theory and practice grounded in richer notions of programme implementation may be an important avenue for future work.

References

Abdul Latif Jameel Poverty Action Lab (J-PAL). 2018. "Teaching at the Right Level to improve learning." J-PAL Evidence to Policy Case Study. Last modified August 2022.

Angrist, N., Djankov, S., Goldberg, P.K., and Patrinos, H.A. (2021) 'Measuring human capital using global learning data.' *Nature* 592, no. 7854: 403–408.

Angrist, N., Bergman, P., and Matsheng, M. (2022) 'Experimental evidence on learning using low-tech when school is out.' *Nature Human Behaviour* 6, no. 7: 941–950.

Angrist, N., Evans, D.K., Filmer, D., Glennerster, R., Rogers, F.H., and Sabarwal, S. (2020) 'How to improve education outcomes most efficiently? A Comparison of 150 interventions using the new Learning-Adjusted Years of Schooling metric.' The World Bank.

Andrews, I., and Kasy, M. (2019) 'Identification of and correction for publication bias.' *American Economic Review* 109, no. 8: 2766–94.

Andrews, I., and Oster, E. (2019) 'A simple approximation for evaluating external validity bias.' *Economics Letters* 178: 58–62.

Bandiera, O., Fischer, G., Prat, A. and Ytsma, E. (2017) 'Do women respond less to performance pay? Building evidence from multiple experiments.' Working paper.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. In American Economic Journal: Economic Policy (Vol. 2, Issue 1, pp. 1–30). American Economic Association.

Banerjee, A., Banerji, R., Berry, J., Duflo, E,, Kannan, H., Mukerji, S., Shotland, M., and Walton, M. (2017a) 'From proof of concept to scalable policies: challenges and solutions, with an application.' *Journal of Economic Perspectives* 31(4), pp.73–102

Banerjee, A.V., Hanna, R., Kreindler, G.E. and Olken, B.A. (2017b) 'Debunking the stereotype of the lazy welfare recipient: Evidence from cash transfer programs.' *The World Bank Research Observer* 32(2), pp.155–184.

Banerjee, A.V., Cole, S., Duflo, E. and Linden, L. (2007) 'Remedying education: Evidence from two randomized experiments in India.' *The Quarterly Journal of Economics* 122, no. 3: 1235–1264.

Banerjee, A., Duflo, E. Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015) 'A multifaceted program causes lasting progress for the very poor: Evidence from six countries.' *Science* 348, no. 6236.

Banerji, R. and Chavan, M. (2016) 'Improving literacy and math instruction at scale in India's primary schools: The case of Pratham's Read India program.' *Journal of Educational Change* 17(4), pp. 453–475.Bauer, M S., Damschroder, L., Hagedorn, H., Smith, J., and Kilbourne, A.M. (2015) 'An intro-duction to implementation science for the non-specialist.' *BMC Psychology* 3, no. 1: 1–12.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., Dorie, V. (2015) 'Weakly informative prior for point estimation of covariance matrices in hierarchical models.' *Journal of Educational and Behavioral Statistics* 40(2), 136–157.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., Liu, J. (2013) ‘A non-degenerate penalized likelihood estimator for variance parameters in multilevel models’. *Psychometrika*, 78, 685–709.

Cunha, F. and Heckman, J. (2007) ‘The technology of skill formation.’ *American Economic Review* 97, no. 2: 31–47.

Deaton, A., and Cartwright, N. (2018) ‘Understanding and misunderstanding randomized controlled trials.’ *Social Science Medicine* 210: 2–21.

Duflo, E., Dupas, P., and Kremer, M. (2011) ‘Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya.’ *American Economic Review* 101, no. 5: 1739–74.

Duflo, E. (2017) ‘The economist as plumber.’ *American Economic Review* 107, no. 5: 1–26. [21] [21] Duflo, A., Kiessel, J., and Lucas, A. (2020) ‘Experimental Evidence on Alternative Policies

to Increase Learning at Scale’. No. w27298. National Bureau of Economic Research.

Evans, D.K., and Popova, A. (2016) ‘What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews.’ The World Bank Research Observer 31, no. 2: 242–270.

Evans, D.K, and Yuan, F. (2020) ‘How big are effect sizes in international education studies?’ *Educational Evaluation and Policy Analysis*: 01623737221079646.

Ganimian, A.J. and Murnane, R.J. (2016) ‘Improving education in developing countries: Lessons from rigorous impact evaluations.’ *Review of Educational Research* 86: 719–755.

Global Education Evidence Advisory Panel (2020) ‘Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are ‘Smart Buys’ for Improving Learning in Low and Middle Income Countries? Recommendations from the Global Education Evidence Advisory Panel.’ The World Bank.

Gechter, M. (2017) ‘Generalizing the results from social experiments: Theory and evidence from Mexico and India.’ Working manuscript, Pennsylvania State University.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2004) *Bayesian Data Analysis: Second Edition*. Taylor Francis.

Gelman, A., Hill, J. (2007) *Data analysis using regression and multilevel hierarchical models*. Cambridge Academic Press.

Gelman, A., and Pardoe, I. (2006) ‘Bayesian measures of explained variance and pooling in multilevel (hierarchical) models.’ *Technometrics*, 48(2), 241–251. Informa UK Limited.

Hanushek, E. A. (1995) ‘Interpreting Recent Research on Schooling in Developing Countries.’ *World Bank Research Observer*, Vol. 10, No. 2.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. In Springer Series in Statistics. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

J-PAL (2013) ‘Improving learning by increasing motivation, targeting instruction, and addressing school governance.’ *J-PAL Policy Insights*.

Kraft, M.A. (2020) ‘Interpreting effect sizes of education interventions.’ *Educational Researcher* 49, no. 4: 241–253.

Kremer, M., Brannen, C., and Glennerster, R. (2013) 'The challenge of education and learning in the developing world.' *Science* 340, no. 6130: 297–300.

Lewbel, A. (2019) 'The identification zoo: Meanings of identification in econometrics.' *Journal of Economic Literature* 57, no. 4: 835–903.

Lockheed, M. and Vespoor, A. (1991) *Improving Primary Education in Developing Countries*. Oxford University Press.

Meager, R. (2019) 'Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments.' *American Economic Journal: Applied Economics* 11, no. 1: 57–91.

Meager, R. (2022) 'Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature.' *American Economic Review* 112, no. 6: 1818–47.

Muralidharan, K., Singh, A. and Ganimian, A.J. (2019) 'Disrupting education? Experimental evidence on technology-aided instruction in India.' *American Economic Review* 109, no. 4: 1426–1460.

Piper, B., Zuilkowski, S.S. and Mugenda, A. (2014) 'Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative.' *International Journal of Educational Development* 37: 11–21.

Pritchett, L., and Sandefur, J. (2015) 'Learning from experiments when context matters.' *American Economic Review* 105, no. 5: 471–75.

Rubin, D.B. (1981) 'Estimation in parallel randomized experiments.' *Journal of Educational Statistics* 6(4), pp. 377–401.

Smale-Jacobse, A.E., Meijer, A., Helms-Lorenz, M., and Maulana, R. (2019) 'Differentiated instruction in secondary education: A systematic review of research evidence.' *Frontiers in Psychology* 10: 2366.

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., Jobse, H., Schmidt, T., and Jimenez, E. (2016) 'The impact of education programmes on learning and school participation in low- and middle-income countries.' *Studies in Regional Science*.

Tomlinson, C.A. (2014) 'The differentiated classroom: Responding to the needs of all learners.' *ASCD.*

UNESCO (2017) 'More Than One-Half of Children and Adolescents Are Not Learning Worldwide.' *UIS Fact Sheet No. 46*.

USAID (2018) 'USAID Education Policy.' Washington, DC.

Vespoor, A. (1989) 'Pathways to Change: Improving the Quality of Education in Developing Countries'. *World Bank Discussion Paper 53*.

Vigneri, M., Masset, E., Clarke, M., Exley, J., Tugwell, P., Welch, V., White, H. (2018) 'Economics and Epidemiology: Two Sides of the Same Coin or Different Currencies for Evaluating Impact?' *CEDIL Inception Paper 10*.

Vivalt, E. (2020) 'How much can we generalize from impact evaluations?' *Journal of the European Economic Association* 18, no. 6: 3045–3089.

World Bank (2018) 'World Development Report 2018: Learning to Realize Education's Promise.' Washington, DC.

Appendix A
Additional Bayesian Models and Results

A.1 Joint Aggregation Model

Since we have access to baseline information about each of the TaRL studies, we can go further than the basic Rubin (1981) model and employ a joint aggregation exercise that leverages this baseline information in order to improve precision and the inferences we can draw. Following Meager (2019), we specify a joint hierarchy on the control group means and treatment effects in each TaRL study, as follows:

$$\hat{\mu}_k \sim N\left(\mu_k, \sigma^2_{\mu_k}\right)$$
$$\hat{\tau}_k \sim N\left(\tau_k, \sigma^2_{\tau_k}\right) \qquad \text{(Equation A.1)}$$
$$\begin{pmatrix}\mu_k \\ \tau_k\end{pmatrix} \sim N\left(\begin{pmatrix}\mu \\ \tau\end{pmatrix}, V\right) \text{ where } V = \begin{bmatrix}\sigma^2_\mu & \sigma_{\tau\mu} \\ \sigma_{\tau\mu} & \sigma^2_\tau\end{bmatrix} \forall k.$$

This joint model incorporates a correlation parameter between the baseline or control group mean and the treatment effects, which can improve precision and estimation overall if such a correlation is present. In other respects it is identical to the classical Rubin (1981) model. This model, developed by Meager (2019), is sometimes referred to as the ‘mu and tau’ model, as in previous literature the effect of a programme was labelled with the Greek letter τ rather than θ (see for example Gelman et al., 2004). This model was shown to substantially improve precision and inference in the microcredit aggregation setting, and is thus worth incorporating into our main analysis in the hope of similar gains to estimation performance (see Meager (2019) for more details).

The results of this model are shown below in Figure 13. They broadly confirm the Rubin (1981) results, which are shown for comparison.

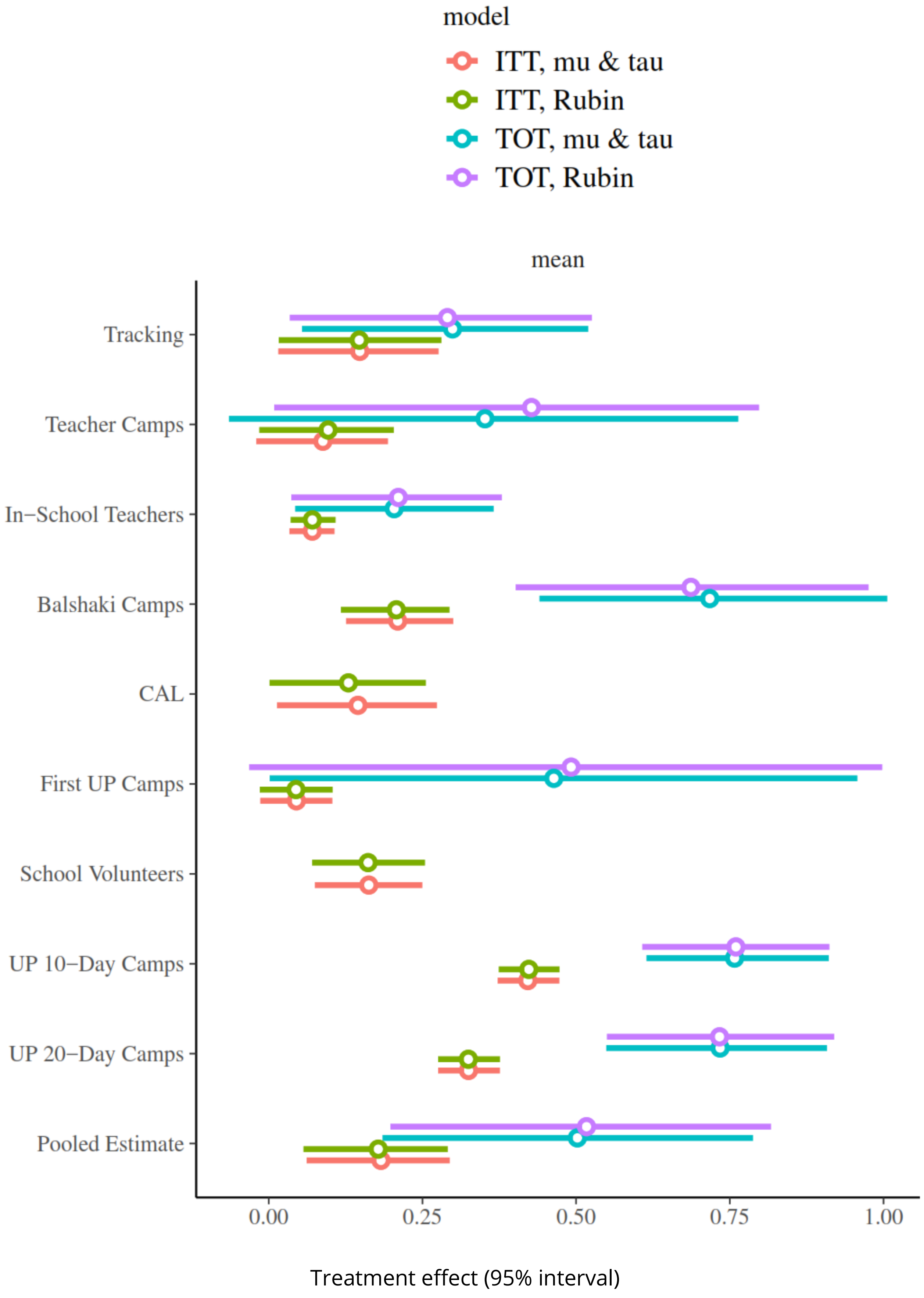


Figure 13: Rubin (1981) model vs joint "mu and tau" model

Study	ITT Pooling Factor
Balshaki Camps	0.10
CAL	0.20
First UP Camps	0.04
Tracking	0.23
Teacher Camps	0.15
School Volunteers	0.11
In-School Teachers	0.02
UP 10-Day Camps	0.03
UP 10-Day Camps	0.03
All (simple average)	0.10

Study	TOT Pooling Factor
Balshaki Camps	0.23
First UP Camps	0.69
Tracking	0.17
Teacher Camps	0.42
In-School Teachers	0.09
UP 10-Day Camps	0.07
UP 20-Day Camps	0.11
All (simple average)	0.25

Figure 14: Individual pooling factors from the simple Bayesian model

A.2 Bayesian Pooling Factors

We now display the pooling factor for each study in the simple Bayesian hierarchical model. Figure 14 shows that both the ITT results and the TOT results are heterogeneous across settings, but the ITT is much more heterogeneous, and the hierarchical models therefore perform much less partial pooling on the ITTs – as the last line of each table shows, only 10% of the variation observed in the original ITT study results was due to sampling variation, while 25% of the variation observed in the original TOT study results is sampling variation.

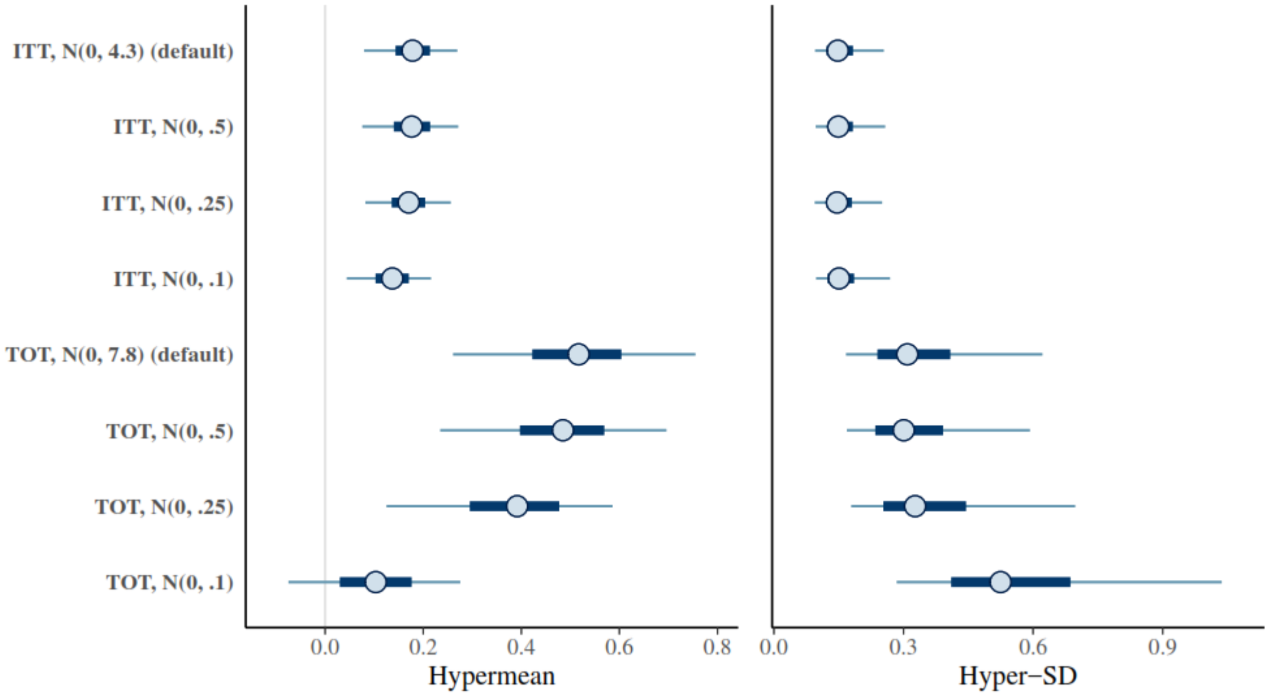


Figure 15: Additional prior robustness checks

A.3 Further Prior Robustness Results

Figure 15 shows that for both the hypermean and hyperSD (heterogeneity in effects across settings) even very strong priors cannot substantially influence the inference on the ITT results. The evidence on the positive impact of targeted instruction at the school level is extremely strong. This contrasts somewhat to the TOT results, which are somewhat more influenced by the priors – this is because, as discussed in our main results section, the TOT effects are estimated with greater uncertainty within each study, and therefore are less able to overcome pessimistic priors. However, the strength of the TOT evidence on a positive average (hypermean) is shown in the compensating pattern in the hyperSD; if the hypermean is forced down closer to zero, the hyperSD is forced upwards to compensate for the fact that we have strong evidence of some effects being large and positive in some studies. Moreover, even with the most pessimistic prior, the Bayesian models report more than 75% chance of a positive effect of targeted instruction. Overall, therefore, the evidence on the positive impact of targeted instruction is strong.

We examine the results of imposing a stronger theory-driven regularisation of the correlation between baseline educational performance and the treatment effect of targeted instruction towards zero. As discussed, the competing theoretical mechanisms suggest we should expect a small correlation; this corresponds to expecting or favouring independence or zero-off-diagonal terms in the variance-covariance matrix. This is implemented via the use of a Lewandowski-Kurowicka-Joe (LKJ) correlation prior distribution on the variance-covariance matrix V from the joint aggregation model described earlier in the appendix. The LKJ distribution is a distribution over the space of correlation matrices, parameterised by a 'concentration parameter' that can take any positive value (see Meager (2019) and Gelman and Hill (2007) for more information). If the concentration parameter is set to be 1, the distribution is uniform over the space of all correlation matrices; if it is larger than 1, it favours independence, expressed by zero off-diagonal terms. The larger the parameter is, the more strongly it favours independence, and thus the more strongly it regularises the correlation in question.

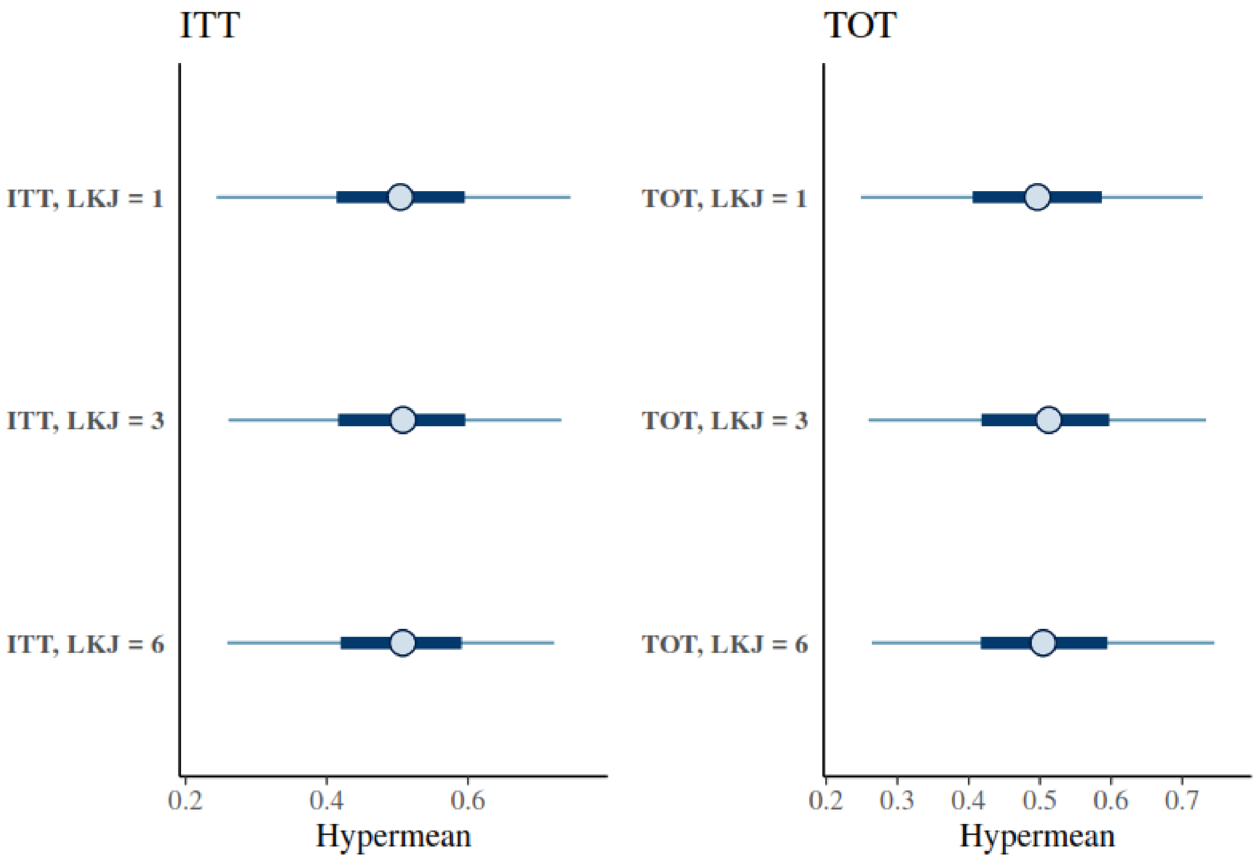


Figure 16: Mu and tau model with different LKJ priors

The graphics in Figure 16 show the results of fitting the joint aggregation model (the ‘mu and tau’ model described above) with an LKJ prior with concentration 1 (the default used in the previous sections), as well as 3 (moderate regularisation) and 6 (strong regularisation). In this case, the stronger priors have no impact at all on the posterior hypermeans, for TOT or ITT. While this is initially surprising, it reflects the fact that the empirical correlation between the baseline and treatment effect in the sample of studies we have is already zero.

These results do not necessarily mean baseline levels of learning do not matter in regard to targeted instruction being effective. Rather, it is possible that the set of studies included are all cases with relatively low baseline levels of learning. Thus, if low baseline levels of learning is a critical condition whereby targeted instruction is needed and effective, for all studies this condition might be met, hence there will be positive effects across the board. This is not mutually exclusive from the insight that the existence of relative differences in baseline learning does not explain heterogeneity in the magnitude of effects conditional on low initial baseline levels of learning.

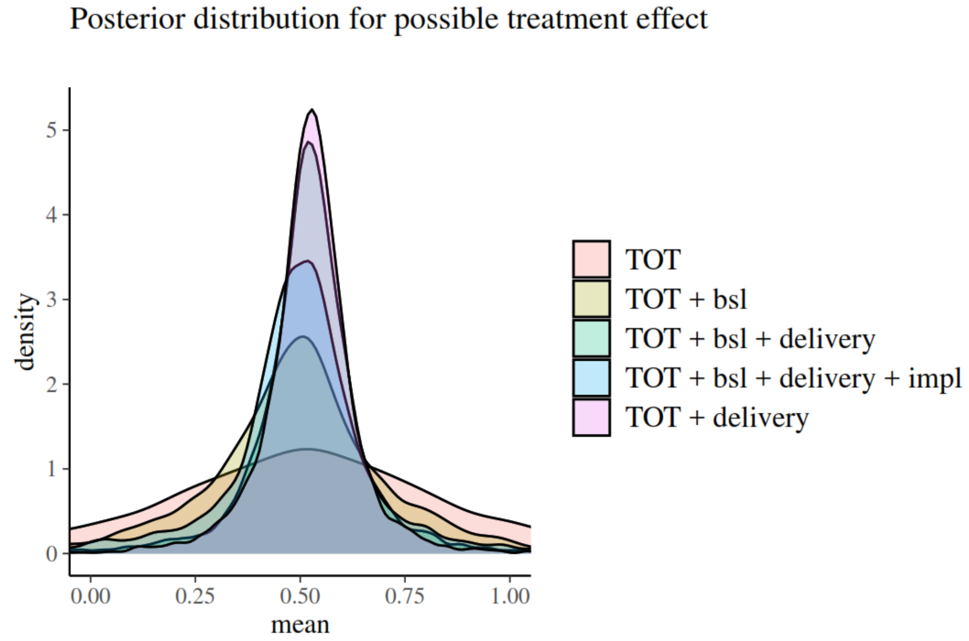


Figure 17: Full posteriors predictive distributions under different models

A.4 Full Posterior Predicted Distribution Graphics

Figure 17 shows the basic model in red, conditioning on baseline in green, on delivery in purple, and on both in blue. As the figure shows, the predicted effect is virtually identical in each case. We examine that same graph for the TOT and here find that the posterior distribution of predicted effect in the next setting is substantially more precise for the model conditioning only on delivery (shown in purple) relative even to conditioning on all possible covariates (blue and green), and certainly much more precise than conditioning on either only baseline or nothing (yellow and red). This result confirms that the remaining heterogeneity in effects here is predicted by delivery mechanism.

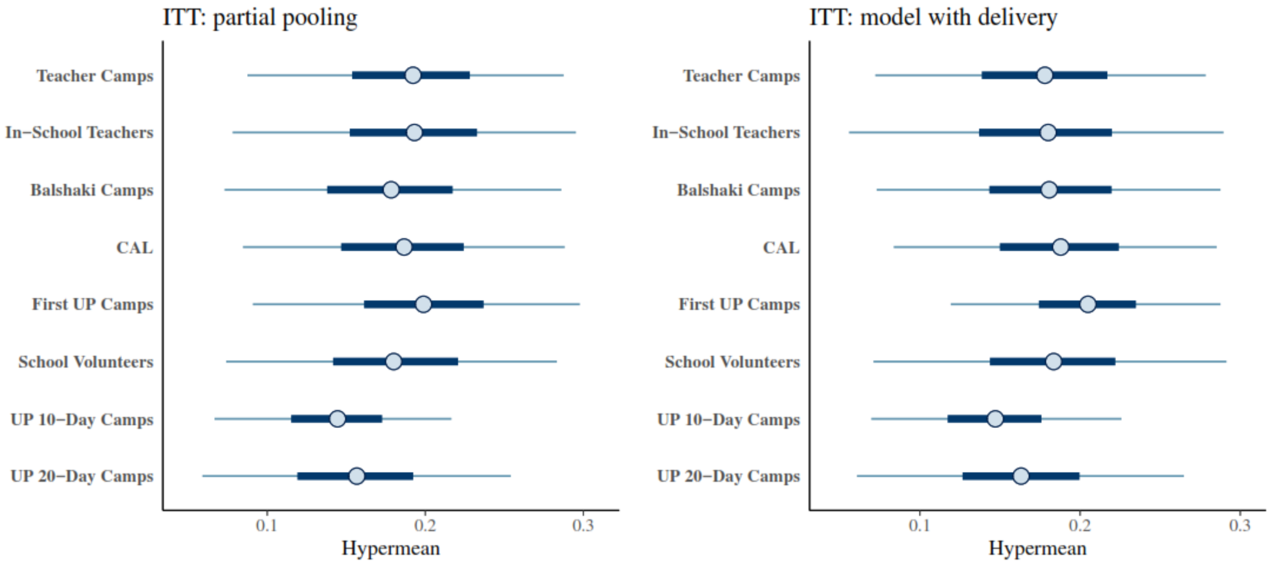


Figure 18: ITT leave-one-out analysis

Appendix B
Evidence Aggregation Robustness Checks

In this appendix we conduct additional robustness checks that help us understand how the inference on the average effects is constructed from the sample of studies at hand, and assess whether the analysis in this paper is vulnerable to classical publication bias.

We first conduct leave-one-out analysis in order to understand the robustness of our main results to omitting any of the studies. This is especially a concern when we have a small number of results in a given literature, as is the case typically for aggregation of RCTs (see for example Meager (2019), where the same robustness check is presented for the microcredit RCT aggregation exercise). We take as our main analytical result of interest the average treatment effect, either in terms of ITT or TOT, across all the studies in the data set. That is, we examine the posterior distribution of the hypermean from the Bayesian hierarchical models and display its sensitivity to leaving out each of the studies in turn. These 'leave-one-out' sensitivity results are shown for the ITT estimates both in the Rubin (1981) partial pooling model and for the model conditioning on the delivery mechanism in Figure 18, where the study indicated in the row label is the study omitted for that run of the model.

The results above show relatively little variation in the posterior distribution of the hypermean when any given study is omitted, with the slight possible exception of the three UP Camps estimates. These three studies each seem to exert more influence than the other studies, although they run in different directions – dropping the first UP Camps tends to increase the hypermean, while dropping the 10- or 20-Day Camps tends to decrease the hypermean. However, in all cases there is substantial overlap in the posterior intervals with the general results, and even for the UP Camps study omission the posterior mean of the hypermean is well within the central 50% credible interval of the other posteriors. This shows relatively strong robustness of the ITT results overall.

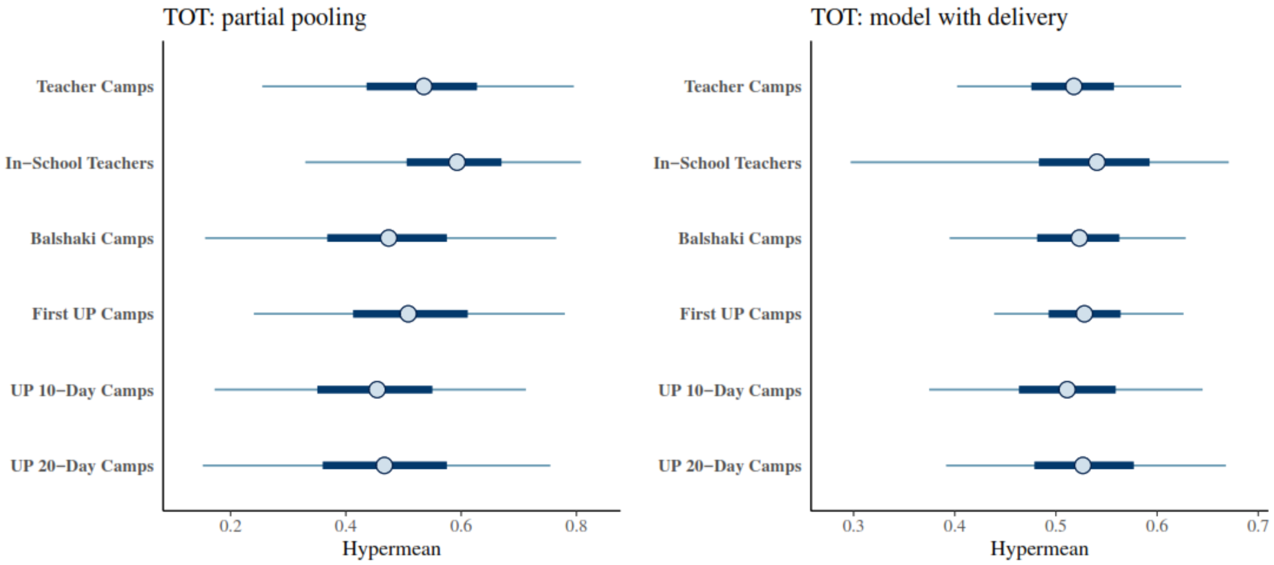


Figure 19: TOT leave-one-out analysis

Figure 19 shows the sensitivity results of leaving out each study in turn for the TOT estimates both in the Rubin (1981) partial pooling model and for the model conditioning on the delivery mechanism below, where again the row label indicates the study omitted from the model run.

The graphs clearly show little variation in the posterior distribution of the hypermean for either the classical Rubin partial pooling model or the meta-regressive model conditioning on delivery. The slight exception is the effect of leaving out the In-School Teachers study in the Rubin model, which has a somewhat more pronounced effect on the hypermean, but this is not present in our preferred meta-regressive specification conditioning on delivery type. This shows the strong robustness of the TOT results conditioning on delivery type to omitting any of the studies, and demonstrates that our insights about the important role of delivery are not based on any single study but rather are borne out across the literature as a whole.

Finally, we explore the potential for publication bias in the targeted instruction literature and the possible impact on our findings. Figure 20 shows the distribution of t-statistics from estimates. We use a test proposed by Andrews and Kasy (2019) where publication bias is probable if we observe a jump in t-statistics right above the 1.96 cut-off, which is a conventional threshold for statistical significance. We do not observe such a jump, and rather observe more studies right under this threshold, as well as t-statistics which are much larger. One potential reason for this distribution is that the sample sizes in this literature are extremely large, limiting the potential for the manipulation of significance thresholds. This ameliorates potential concerns about publication bias being responsible for the overall positive findings on the impact of the targeted instruction intervention, as we find no evidence of any manipulation of t-statistics in our set of studies.

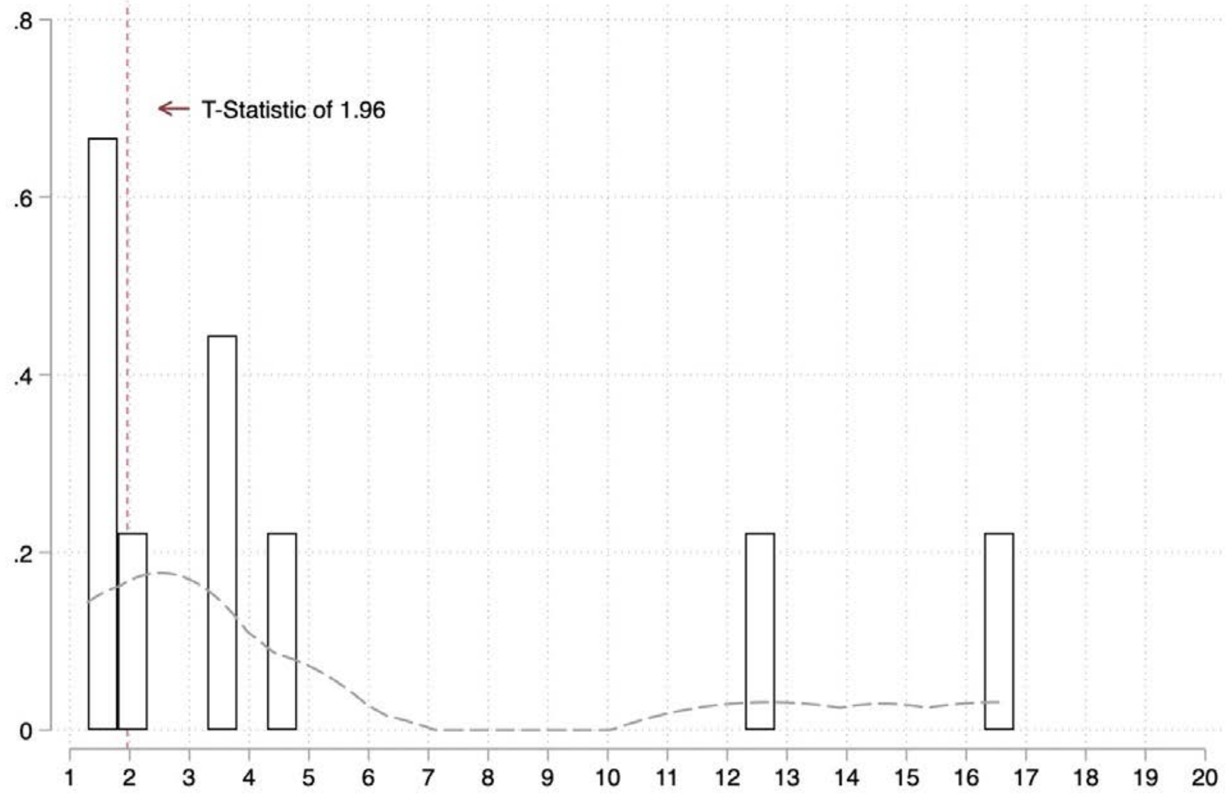


Figure 20: Distribution of t-statistics in our sample

Centre of Excellence for Development Impact and Learning

@CEDILProgramme

https://cedilprogramme.org

Centre of Excellence for Development Impact