ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-802

Implementation Matters: Generalizing Treatment Effects in Education

Noam Angrist
University of Oxford, Youth Impact

Rachael Meager
London School of Economics

Targeted instruction is one of the most effective educational interventions in low- and middle-income countries, yet reported impacts vary by an order of magnitude. We study this variation by aggregating evidence from prior randomized trials across five contexts, and use the results to inform a new randomized trial. We find two factors explain most of the heterogeneity in effects across contexts: the degree of implementation (intention-to-treat or treatment-on-the-treated) and program delivery model (teachers or volunteers). Accounting for these implementation factors yields high generalizability, with similar effect sizes across studies. Thus, reporting treatment-on-the-treated effects, a practice which remains limited, can enhance external validity. We also introduce a new Bayesian framework to formally incorporate implementation metrics into evidence aggregation. Results show targeted instruction delivers average learning gains of 0.42 SD when taken up and 0.85 SD when implemented with high fidelity. To investigate how implementation can be improved in future settings, we run a new randomized trial of a targeted instruction program in Botswana. Results demonstrate that implementation can be improved in the context of a scaling program with large causal effects on learning. While research on implementation has been limited to date, our findings and framework reveal its importance for impact evaluation and generalizability.

VERSION: June 2023

Suggested citation: Angrist, Noam, and Rachael Meager. (2023). Implementation Matters: Generalizing Treatment Effects in Education. (EdWorkingPaper: 23-802). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ysyd-t435

Implementation Matters: Generalizing Treatment Effects in Education

Noam Angrist and Rachael Meager*

Version: June 11, 2023

Abstract

Targeted instruction is one of the most effective educational interventions in low- and middle-income countries, yet reported impacts vary by an order of magnitude. We study this variation by aggregating evidence from prior randomized trials across five contexts, and use the results to inform a new randomized trial. We find two factors explain most of the heterogeneity in effects across contexts: the degree of implementation (intention-to-treat or treatment-on-the-treated) and program delivery model (teachers or volunteers). Accounting for these implementation factors yields high generalizability, with similar effect sizes across studies. Thus, reporting treatment-on-the-treated effects, a practice which remains limited, can enhance external validity. We also introduce a new Bayesian framework to formally incorporate implementation metrics into evidence aggregation. Results show targeted instruction delivers average learning gains of 0.42 SD when taken up and 0.85 SD when implemented with high fidelity. To investigate how implementation can be improved in future settings, we run a new randomized trial of a targeted instruction program in Botswana. Results demonstrate that implementation can be improved in the context of a scaling program with large causal effects on learning. While research on implementation has been limited to date, our findings and framework reveal its importance for impact evaluation and generalizability.

JEL Codes: I2, I25, C11
Keywords: Education, Development, External Validity, Bayesian Analysis

*Angrist: University of Oxford, Youth Impact; Meager: London School of Economics. We thank Witold Wiecek and Veerangna Kohli for valuable research contributions as well as Colin Crossley and Patience Derera for excellent grant and program administration. We thank Tendekai Mukoyi, Thato Letsomo, and Moitshepi Matsheng for their programmatic and organizational leadership. Thank you to Stefan Dercon, Clare Leaver, Adrienne Lucas, Eduardo Masset, and Abhijeet Singh for detailed comments. We thank Josh Angrist, CJ Angrist, Abhijit Banerjee, Claire Cullen, Emily Cupito, Andy de Barros, Esther Duflo, Rachel Glennerster, Macartan Humphreys, Harini Kannan, Dean Karlan, Michael Kremer, David McKenzie, Ashleigh Morrell, Shobini Mukherji, Jennifer Opare-Kumi, and Lant Pritchett for 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 and conducted randomized A/B tests on a scaling program. Noam Angrist is a co-founder of the NGO 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 the Center of Excellence for Development Impact and Learning (CEDIL) which is supported by the United Kingdom Foreign Commonwealth and Development Office (FCDO).

1 Introduction

617 million young people worldwide are in school but unable to read fluently or perform simple numerical operations. These learning deficits are particularly acute in developing countries (UNESCO 2017; World Bank 2018; Angrist et al. 2021). Although many policies designed to address the learning crisis have yielded disappointing results, programs which target educational instruction to a child's learning level have improved learning in a variety of contexts.[1] Randomized trials show consistently positive impacts in India, Kenya, and Ghana, receiving significant attention in academic and policy circles (Banerjee et al. 2007; Banerjee et al. 2010; Duflo, Dupas and Kremer 2011; Banerjee et al. 2017; Duflo, Kiessel, and Lucas 2020).[2] A recent high-profile report highlighted targeted instruction as a cost-effective approach to address the global learning crisis (Global Education Evidence Advisory Panel 2020). However, while effects are consistently positive, they range from 0.07 to 0.78 standard deviations – an order of magnitude difference.[3] Systematic analysis of this variation could reveal important information on how generalizable effects might be, and uncover factors that yield the largest frontier effects in the literature. This is especially important as targeted instruction is adapted across contexts with multiple ambitious scale-up efforts underway.

In this paper, we first assess the generalizability of targeted instruction by aggregating evidence across prior randomized trials. We then use the results to inform a new randomized trial, optimizing delivery of a targeted instruction scale-up in Botswana. For our aggregation, we consider data across 8 study arms covering nearly 75,000 students across 5 contexts. We collect data on effect sizes as well as contextual covariates such as baseline learning, geographical context, sample size, year, and implementation delivery model (teachers or external volunteers). We also consider data on program implementation first using the notion of "takeup", measured by attendance or presence of classroom materials. Second, we consider the "fidelity" of implementation, measured by adherence to core program principles, such as whether instruction is targeted and students are grouped as expected. Targeted instruction is an ideal setting in which to study the different role of these two aspects of implementation, as both vary widely across studies: takeup ranges from 8% to 90%, and fidelity from 23% to 83%.

Given heterogeneity in both program features and reported effects, careful attention to evidence aggregation methodology is required. We provide results from both standard Frequentist random-effects meta-analysis and a series of Bayesian hierarchical models, including meta-regression models which formally incorporate data on program features. We also report several metrics of generalizability: first, the frequentist I-squared metric. A low I-squared indicates that most observed variation in effects is sampling variation, rather than true treatment effect variation, indicating high generalizability. We also report the Bayesian hyper-standard-deviation which measures the standard deviation in true effects, and the Bayesian posterior predictive

[1] Targeted instruction groups students in classrooms and tailors instruction to each students' actual learning level rather than to an average expected learning level determined by a one-size-fits-all grade-level curriculum. A specific model of this approach called "Teaching at the Right Level" has been pioneered by Pratham, a large education 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] It is important to note that these reported effects are all substantial in a context where a 0.10 standard deviation effect size is considered large (Kraft 2020; Evans and Yuan 2020).

distribution which captures all uncertainty about the predicted effect in the next hypothetical study setting. Given a sample of 9 study arms, which is large for evidence synthesis, but small for frequentist statistical approaches relying on large-sample properties, we prefer the Bayesian approach for evidence aggregation in our setting.

While many factors might explain effect size differences across contexts, such as geography, baseline learning, or the scale of a given study, results show that most of the heterogeneity in reported effects can be explained by two implementation factors: implementation delivery model (teachers or volunteers) and the degree of implementation (intention-to-treat or treatment-on-the-treated effects). The frequentist random-effects meta-analysis finds that intention-to-treat (ITT) effects for teachers are moderate (0.07 SDs on average) and highly generalizable (I-squared of 0.01%), whereas volunteers have large average effects (0.24 SDs) with high variation (I-squared of 95.6%). When accounting for implementation, treatment-on-treated (TOT) effects for teachers are three times larger (0.21 SD) and are generalizable (I-squared of 0.00%). Similarly, effects for volunteers are three times larger, with an average effect of 0.76 SDs. Most strikingly, effects now converge almost fully, with TOT effects showing high generalizability with an I-squared of 0.00%. This result reveals that much of the original heterogeneity in volunteer estimates was due to variation in implementation. Thus, conditional on implementation factors, the effects of targeted instruction appear large and highly generalizable across studies.

The Bayesian analysis upholds these patterns, although results are somewhat tempered.[4] The TOT effect is still much larger than the ITT on average, particularly for volunteer delivery, and highly generalizable across settings. Moreover, individual studies' TOT estimates see large gains in precision due to partial pooling, combining information from low and high implementation settings.[5] Bayesian meta-regression confirms that implementation takeup 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. The evidence on the positive impact of targeted instruction is clear even when we impose strong null priors (a form of Ridge regularization), which 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 program implementation predict the largest effects in the literature as well as generalizability of effects across settings.

We introduce a new framework which formally incorporates data on implementation into the evidence aggregation process. The analysis thus far has considered implementation and other covariates to be fixed numbers. Yet, implementation features are random variables about which we are uncertain and which may be correlated. To formally account for this, a Bayesian approach permits us to jointly account for uncertainty in effects and uncertainty in implementation. Our model offers several theoretical results, namely that neither treatment effects nor their variation across settings are identified in the absence of information on implementation, or if implementation is poor. For intuition, consider the case of null treatment effects. Null effects could be due to an ineffective program or an effective program which was never implemented. Without information on implementation, a null result may be misattributed to a treatment effect when in fact

[4]This is due to accounting for joint uncertainty in effect averages and variation (Meager 2019).

[5]For example, First UP Camps becomes statistically significant by conventional Frequentist standards.

it is null implementation.[6] Simulations show that the simplest version of the model performs well even with small samples. We apply our framework to our data. Results show that targeted instruction offers 0.42 standard deviation improvements in learning on average when fully taken up, and 0.85 standard deviation gains when implemented with high fidelity, consistent with the upper range of effects in the literature.

Our evidence aggregation and model establish the importance of implementation, motivating research into concrete ways to increase program takeup and fidelity. We conduct a randomized trial optimizing fidelity for a targeted instruction intervention that is being scaled up in Botswana. As of 2022, over 20 percent of primary schools in the country were reached through a partnership between the Ministry of Education and Youth Impact, one of the largest NGOs in the country. We randomly vary implementation fidelity – achieved via more detailed learning assessments and grouping of students relative to standard implementation, enabling even more targeted instruction – in a subsample of 52 classes and over 1000 students in 4 regions. We find that improved fidelity increases the program's impact by up to 0.22 standard deviations (SD). 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 favorable settings yield both high implementation and large effects; rather, improving implementation directly improves program results holding all else equal. Overall, we find that implementation factors are decisive in both the size and generalizability of a program's impact, and that implementation can be further optimized in the context of a scaling program.

Our findings contribute to a literature on education in low- and middle- income countries. Improving learning outcomes is difficult, with decades of stagnant learning outcomes, despite increasing enrollment in school (Pritchett 2013; Angrist et al. 2021). Moreover, input-focused interventions which simply provide more resources, such as provision of textbooks or computer hardware only, have been found to rarely improve learning (Glewwe, Kremer, and Moulin 2009; Kremer, Brannen, and Glennerster 2013; Beuermann et al. 2015). 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. 2017; Muralidharan, Singh, and Ganimian 2019; Duflo, Kiessel, and Lucas 2020). Our results are consistent with this emerging view and show that this insight generalizes across heterogeneous contexts. In contrast with evidence aggregations in other sectors, such as microcredit, which show small or null effects (Meager 2019), we find that targeted instruction has large and generalizable effects.

We also contribute to the literature on external validity and advance the practice of evidence synthesis. Although systematic evidence aggregation is rare in economics, researchers are increasingly engaging in evidence synthesis across contexts (Banerjee et al 2017b, Andrews and Oster, 2019; Bando, Näslund-Hadley, and Gertler 2019; Vivalt 2020; Bandiera et al. 2021; Meager 2022; Gechter 2023). We contribute methodologically to this literature along a few dimensions. We are one of the first to directly combine evidence aggregation of prior studies

[6] A full exposition is provided in the paper. Here we summarize a key component. We define Realized Treatment Effects (RTE) as equal to Latent Treatment Effects (LTE), θ_j, multiplied by an implementation factor, $m_j \in [0, 1]$ such that $RTE = m_j * \theta_j$. A program 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 can not be logically used to infer a null latent treatment effect θ_j. In other words, the underlying effect θ_j is not identified from the data.

with a new randomized trial, linking experimental and synthesis methods. Second, we show that accounting for implementation through treatment-on-the-treated estimates enables better generalization of effects across contexts. TOT estimation has well-developed frameworks (Imbens and Angrist 1994), but remains infrequent in practice. Out of a set of papers in development from 2019 and 2022, for example, all report ITT effects but only 12.1 percent of education RCTs reported TOT effects or implementation metrics.[7] Our results show that in order to achieve external validity, reporting and aggregating TOT estimates should become common practice. Third, while internal validity questions have several frameworks, fewer formal frameworks exist for external validity. Informed by our results, we introduce a new framework for external validity. We formalize the essential role of implementation information, which we refer to as *m-factors*, in identification of treatment effects and for generalizability. We further present a novel Bayesian model incorporating implementation into the evidence aggregation process.

Finally, we advance a nascent literature on implementation science in education. Implementation science is best developed in health, where it is viewed as a largely qualitative concept (Bauer et al. 2015). This paper demonstrates how to quantitatively account for implementation. Our results also show that implementation is the decisive dimension in generalizing results across contexts, motivating a research agenda on the details of effective implementation, consistent with the notion of "the economist as plumber" (Duflo 2017). Given that targeted instruction yields 0.85 SD when delivered with high fidelity – 10-fold higher than the typical education intervention – research on better implementing known productive interventions, such as targeted instruction, can be higher return than discovery of new interventions. Our trial in Botswana provides a concrete example of the return to studying implementation fidelity.

The results in this paper have significant implications for policy. Targeted instruction has the potential to help address the global learning crisis and is on track to reach over 60 million children in South Asia and sub-Saharan Africa by 2025 (J-PAL 2022).[8] High-profile scale up examples include Zambia, where the government has already scaled up to over 3,000 schools, and Nigeria where targeted instruction is being delivered in over 5 states; in addition, in Botswana the government has signed a 9-year Memorandum of Understanding to scale-up nationally. This paper estimates the generalizability of targeted instruction and identifies factors that mediate the largest effects in the literature, informing adaption and scale-up in new contexts.

The rest of the paper is organised as follows. Section 2 describes the intervention and context, Section 3 describes the data, and Section 4 explains the evidence aggregation approach. Results are presented in Section 5. In Section 6 we introduce a new framework to include implementation metrics in evidence aggregation; we further formalize the role of implementation in identification of treatment effects and for generalizability. Section 7 includes results from a new randomized trial in Botswana optimizing implementation of a scaling program, and Section 8 concludes.

[7] We conduct a review to identify how frequent the practice of accounting for implementation is in program evaluation. Out of 4,000 papers in development from 2019 to 2022, nearly 25 percent were RCTs; of those that were RCTs in education, only 12.1 percent reported implementation metrics or TOT estimates. We include papers captured in the 3ie database, the Top 5 economic journals (American Economic Review, Quarterly Journal of Economics, Econometrica, Journal of Political Economy, Review of Economic Studies), top-tier general interest journals (Review of Economics and Statistics, Economic Journal, Journal of the European Economic Association, all American Economic Journal AEJ journals), and a top field journal (the Journal of Development Economics).

[8] See J-PAL website: https://www.povertyactionlab.org/case-study/teaching-right-level-improve-learning

2 Educational Intervention and Context

Educational enrollments have increased worldwide to above 90% in all regions, yet learning progress has been much more limited. International education agencies have called this phenomenon a "learning crisis" (World Bank 2018). The 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 129 billion USD in lost social welfare (UNESCO 2017).

A combination of factors contributes to the learning crisis, including curricula targeted mostly to advanced students, rote learning, and automatic promotion regardless of learning achieved in prior grades (Banerji and Chavan 2016). Many education interventions have focused on providing inputs to improve learning, such as textbooks, computers, cash transfers, reducing class size, or increasing teacher salaries. However, decades of randomized trials show input-focused initiatives rarely improve learning (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 randomized trials to dramatically improve learning across multiple contexts including India (Banerjee et al. 2017), Kenya (Duflo, Dupas and Kremer 2011), and Ghana (Duflo, Kiessel, and Lucas 2020). Targeted instruction involves regrouping students by their learning level (e.g., addition, subtraction) rather than using grade-level grouping determined by rigid curricula. Most education systems are organized to teach a one-size-fits-all curriculum by grade. However, there is often substantial heterogeneity in student learning level in each grade, with most students well below grade-level expectations. 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. Yet, if only 10 percent of the class knows division, 90 percent of the class is being left behind. If a child cannot recognize or add numbers, they will not be able to learn division. Targeting instruction involves regrouping students by learning proficiency rather than by grade. Instead of using mass education that reaches a few, this approach uses customized and engaging teaching and learning that is targeted to the learning level of the child.

A specific model of targeted instruction approaches is called "Teaching at the Right Level" (TaRL), developed by Pratham, one of the largest education NGOs in India. Targeted instruction approaches have been shown to consistently improve learning outcomes for children across diverse contexts.[9] But as yet no systematic meta-analysis across studies has been conducted. A systematic analysis could help identify factors which drive heterogeneity and predict the highest frontier effects in the literature. Making progress on this question has significant implications as targeted instruction approaches are 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 actively engaged in targeted instruction scale-up efforts (see Figure A1).

[9] 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 analyze microdata across a series of existing clustered randomized controlled trials conducted over the last two decades across India and Kenya.[10] In total, these trials represent eight geography-treatments. The total sample across the studies is nearly 75,000 students. We start our analysis with studies recognized as consistent with the targeted instruction model by both the original evaluators and implementers.[11] Moreover, we access the microdata for all these studies, allowing us to replicate original results as well as enhance our ability to conduct a comprehensive evidence aggregation exercise capturing study-level covariates and program features. We focus our analysis on randomized controlled trials to ensure we aggregate causal effects.

Table 1 lists included studies and highlights key sample characteristics for each. We consider the relevant level of observation to be the study-treatment-geography. States in India have a population on the scale of most countries, and are highly heterogeneous, so we include states as a geographical unit in India; moreover, the trials in India often stratify or randomize within a state.

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	10000
Banerjee et al. (2010)	Uttar Pradesh, India	First UP Camps	Volunteer	9442
Duflo et al (2011)	Kenya	Tracking	Teachers	6000
Banerjee et al. (2017)	Bihar, India	School Volunteers	Volunteer	3325
Banerjee et al. (2017)	Bihar, India	Teacher Camps	Teachers	2474
Banerjee et al. (2017)	Uttar Pradesh, India	UP 10-day Camps	Volunteer	17266
Banerjee et al. (2017)	Uttar Pradesh, India	UP 20-day Camps	Volunteer	13054
Banerjee et al. (2017)	Haryana, India	In-school Teachers	Teachers	11966
Total	5	8	-	73527

Outcome Data and Measurement. We access the microdata from each study to produce new standardized outcome variables. In all studies, the central outcome is a measure of learning basic numeracy and literacy skills. Most studies use an assessment similar to the Annual Status of Education Report (ASER) test. Figure A2 shows examples of ASER assessments for literacy and numeracy. Our primary outcome is an average of numeracy and literacy.

The ASER test is a validated learning measure which tests competencies used across 14 countries and is consistently used in the education literature (Banerjee et al. 2017). In numeracy,

[10] In the future, we hope to incorporate forthcoming results from studies in Ghana.

[11] For example, we exclude some treatment arms in Banerjee et al. (2017) which only provided generic materials, but did not provided targeted instruction support. We also exclude some treatments where targeted instruction did not occur in practice such as in Uttarakhand as verified in carefully collected monitoring data.

questions include number recognition, addition, subtraction, multiplication, and division. In literacy, 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 standardize each score relative to the standard deviation within a state/country-treatment unit. We derive learning gains over the course of one year to compare outcomes on a consistent time horizon. Given our underlying assessments measure a similar outcome and most use a similar test, comparability of outcomes across contexts is high.

One of the advantages of the targeted instruction intervention is that the outcomes are similar across intervention settings. In many cases, meta-analyses rely on outcomes which can vary substantially, derived using entirely different surveys and definitions of outcomes, such as in the case of microcredit (Meager 2019, Vivalt 2020, Pritchett and Sandefur 2015). Our relatively uniform outcome data is well suited to aggregation and offers a substantial improvement in the comparability of treatment effects across contexts. Throughout the paper, we use standard deviation (SD) units, a common unit in the education literature.[12]

Implementation data. Most studies report intention-to-treat (ITT) effects using random assignment to estimate treatment effects. We replicate intention-to-treat effects. We also estimate treatment-on-the-treated (TOT) effects for those who actually received the program. In some cases TOT effects were originally reported, but not in all cases, so we calculate new estimates for all studies for consistency.[13] Capturing the degree of implementation via take-up is likely to be central to understanding both the average impact and the generalizability of the evidence, since in some studies the degree of implementation is over 80 percent while in other studies it is around 10 percent. It is further noteworthy that in many cases implementation was very high. This reveals that while implementation can vary, it can also reach near-complete levels, increasing the relevance of understanding effects under full take up.

We consider four measures to capture distinct aspects of program implementation: (a) teacher attendance (b) student attendance (c) materials usage and (d) whether students were grouped by learning level. We define takeup as a combination of the first three measures (a)-(c) available across nearly all studies. The last measure goes beyond takeup to capture implementation fidelity – how targeted the instruction was – however this measure is only available in three intervention arms. We incorporate this information into our Bayesian analysis, which is best suited to small-sample statistical inference.

Additional data and covariates. We standardize and incorporate a series of additional data and covariates likely to mediate effects across studies. These include: geography (country or

[12]The practice of using SD units is widespread. However, comparable effects on the raw scale may diverge in SD units, or vice versa; this may be an important topic for future work.

[13]We access the microdata from original studies to quantify the degree of implementation. We have takeup data for all studies, and we have fidelity data for a subset of studies.

state), implementation delivery model (teacher or volunteer), year of intervention, sample size, baseline learning levels, and the degree of implementation.[14]

Replication. We replicate original results prior to conducting a new meta-analysis. We find broadly consistent results with original reported estimates. No signs change and findings remain robust. Average differences across all studies are less than 0.1 standard deviations. In a few rare cases, the magnitude of estimates differ from original estimates, but only slightly. Reasons for this variation include use of a midline rather than an endline assessment, different construction of standard deviations, and our primary measure being an average of scores across both subjects. In addition, in our replication we do not include control variables which can have minor effects on final estimates.

4 Evidence Synthesis

We systematically analyze the variation in treatment effects of targeted instruction and conduct a generalizability analysis. We use the term generalizability, capturing whether evidence translates to broader populations, in contrast with transferability, which refers to extrapolation from one specific setting to another. The term external validity, most common among economists, is sometimes used to refer to both concepts. We conduct various types of meta-analyses. An advantage of meta-analyses is quantitative synthesis of evidence. Moreover, systematic synthesis generates statistics which can be used to gauge average effects sizes across contexts and generalizability, such as the I-squared statistic. The I-squared metric measures the percentage of total variation which is genuine variation in treatment effects rather than sampling variation. A low I-squared indicates that most heterogeneity is due to sampling variation, rather than true treatment effect variation, indicating high generalizability. We also report the Bayesian hypervariance which measures the variation in effects, and the posterior predictive distribution which captures uncertainty about the predicted effect in the next hypothetical study setting.

In some disciplines meta-analyses are seen as the next tier in evidence strength after randomized controlled trials, enabling systematic aggregation of internally valid studies across studies and contexts. However, others argue that meta-analyses are atheoretical and often compute average effects 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 inform our meta-analysis with theory and qualitative expertise. Bayesian synthesis lends itself particularly well to this approach, especially since one can 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.

[14] 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.

4.1 The Bayesian Hierarchical Approach

Aggregating evidence from different settings requires joint estimation of average effects and heterogeneity in effects across studies. The statistical challenge is to separate genuine heterogeneity in effects from sampling variation and simultaneously use this variation to inform the uncertainty on 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 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).[15]

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. (2021) and the Cochrane Handbook version 5.1 section 16.8. This approach provides an initial estimate of the degree of generalizability. We discuss our modelling approach, followed by meta-regression, and then statistics to measure generalizability.

4.2 The Hierarchical Modeling Approach

We start with 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, since they use a consistent estimator and invoke the Central Limit Theorem. These assumptions underlie the computation of confidence intervals and p-values in frequentist research papers, imposing no additional structure on the original papers' analyses. The hierarchical component of the model 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 which governs their dispersion around the true effect. The resulting hierarchical likelihood is written as follows:

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

This parametric model is more general than it appears: for example, if the hyperSD is set to 0 this model nests classical Frequentist fixed-effects meta-analysis. The model can be estimated in a frequentist manner, although it can be easier for this model to be estimated using Bayesian methods. The model's performance has been extensively discussed in Gelman et al. (2004) and has 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 percent of the time. The key assumption embedded in the model is that of exchangeability between the

[15]The Bayesian hierarchical framework also permits multiple comparisons, automatically adjusting for multiple testing problems since marginalization of the joint posterior appropriately conditions on all the evidence available in the sample and priors.

effects being studied, which is a weaker form of the classical i.i.d assumption, and plausible in meta-analytic settings (Meager 2019).

To estimate this likelihood model in a fully Bayesian manner it is necessary to include a prior – that is, adding a prior distribution to the hyperparameters. Following Gelman et al. (2004) and Meager (2019), we use weakly informative priors as a default approach: this imposes some structure without unduly influencing the posterior results. For the hypermean, we tend to center 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. Priors can improve overall estimation by making a favorable bias-variance tradeoff even when the prior information is incorrect. It is also possible to encode more substantive information in priors, such as basing priors on economic theory or contextual knowledge. We conduct analysis with informative priors in Appendix A5 and A6 and in Section F.1. This provides both results of substantive interest as well as robustness checks on the strength of patterns in the data in the face of strong priors.

4.3 Meta-regression within the Hierarchical Framework

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

$$\begin{aligned} \hat{\theta}_j &\sim N(\theta_j, \hat{se}_j) \\ \theta_j &\sim N(X_j\beta, \sigma_\theta^2) \end{aligned} \tag{4.2}$$

While the model is fully parametric, it can be generalized: 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. This is because the kernel of the Gaussian likelihood corresponds to the Ordinary Least Squares objective function.

One important note about this model is that in the context of few studies, say J < 15, and many covariates of interest, say K > 3, then the estimation of the regression coefficients β becomes very challenging and noisy. Moreover, the problem may be masked due to the risk of overfitting. However this issue can be addressed via the use of a Machine Learning technique known as regularization (Hastie et al 2009). Regularization involves the incorporation of a penalty function to prevent an estimation procedure from freely wandering around the parameter space. Classic 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.[16] We use this penalty throughout our hierarchical meta-regression.

4.4 Assessing Heterogeneity and Generalizability

There are several approaches within the Bayesian framework to assessing heterogeneity in effects and generalizability of results. The hyperSD parameters capture the population variation in effects and deserve particular attention. We report the results on the hyperSD throughout, which is usually treated as the fundamental parameter in Bayesian hierarchical models (Gelman et al. 2004). However, it is challenging 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 to sampling variation that causes estimates to vary more than the true effects. This is the reciprocal metric to the 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, this indicates true treatment effect variation is higher than sampling variation – that is, the heterogeneity across settings dominates the uncertainty within settings. This makes extrapolation across settings challenging and suggests low generalizability. In this case, pooling of effects across studies is low. Conversely, when I-squared is low, sampling variation is larger than true treatment effect heterogeneity across settings. This corresponds to relatively high external validity, and the Bayesian hierarchical model will have higher pooling factors and perform more "partial pooling" of effects.

Posterior predictive distributions provide another metric to assess heterogeneity in effects across settings. These distributions capture the uncertainty about the hypothetical treatment effect in the next study. In the Rubin (1981) model, when defining the posterior distribution of these hyperparameters as $F(\theta, \sigma_\theta)$, the posterior predictive distribution for the next effects is:

$$\theta_{J+1} \sim N(\theta, \sigma_\theta^2 | F(\theta, \sigma_\theta^2). \tag{4.3}$$

If one uses an aggregation approach that does not explicitly measure heterogeneity in effects across programs 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. Posterior predictive distributions capture how heterogeneity across settings impedes or enables our ability to extrapolate evidence to the future targeted instruction intervention. This is natural metric of generalizability, linking the present evidence base to future settings and is presented in Figure A3.

[16] As discussed in Hastie et al (2009), in Bayesian analysis a Gaussian prior on the regression coefficients centered at zero is analytically identical to a Frequentist Ridge Regression penalty.

5 Evidence Aggregation Results

5.1 Frequentist Random-Effects Results

We start with a Frequentist random effects aggregation, contextualizing the Bayesian aggregation results by providing results without any formal incorporation of theory or priors. First, we aggregate the evidence on the Intention to Treat (ITT) effects, shown in Figure 1. We find that intention to treat effects for interventions delivered by teachers have an average effect of 0.07 standard deviations. 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 generalizable across the programs in our data set. Second, we observe that volunteer delivery is on average three times as effective as teacher delivery with a 0.24 standard deviation effect. However, the volunteer results are highly heterogeneous with an I-squared of 95.6 percent.

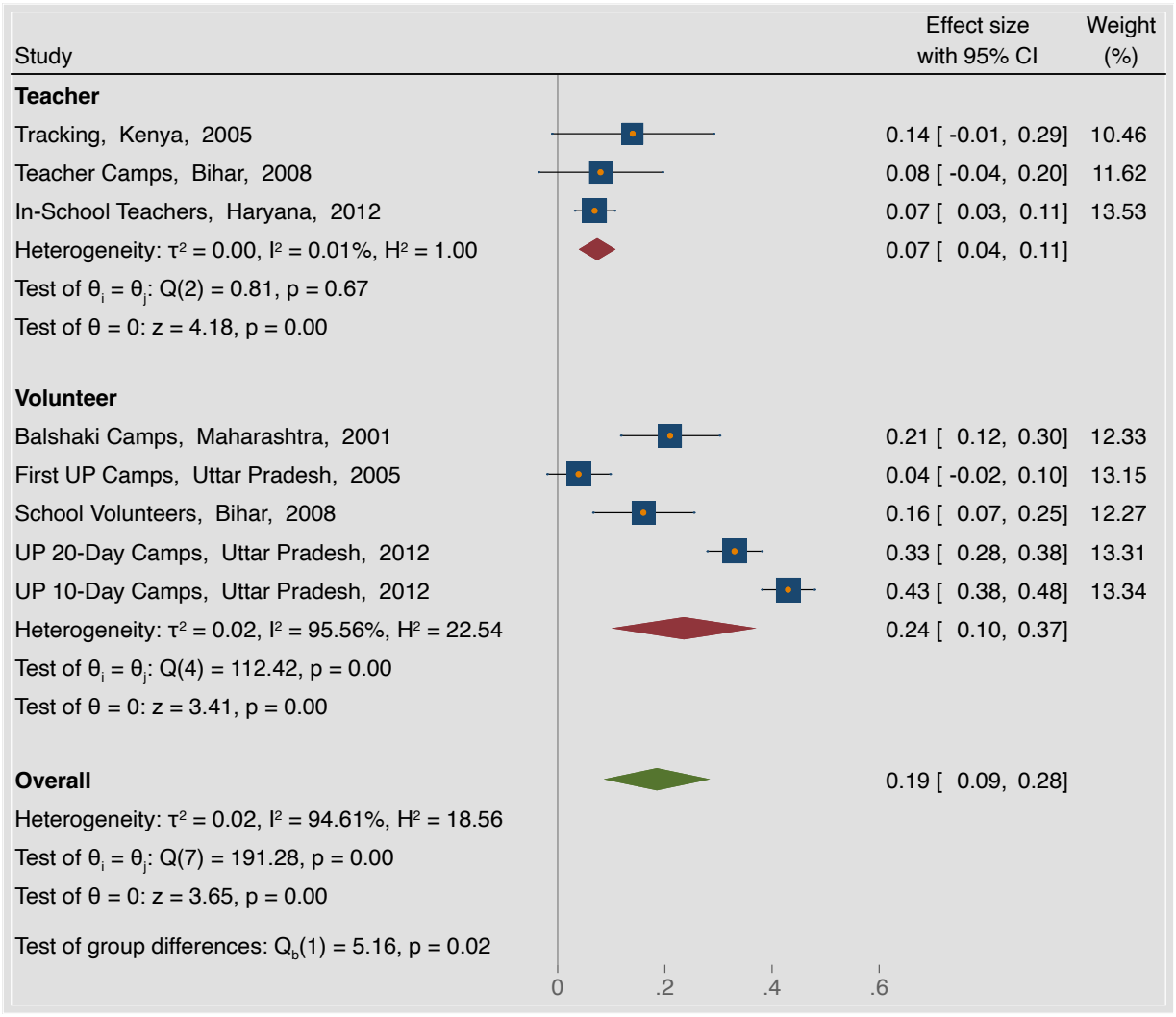

Figure 1: Frequentist Random Effects Meta-Analysis of Intention-to-Treat Effects.

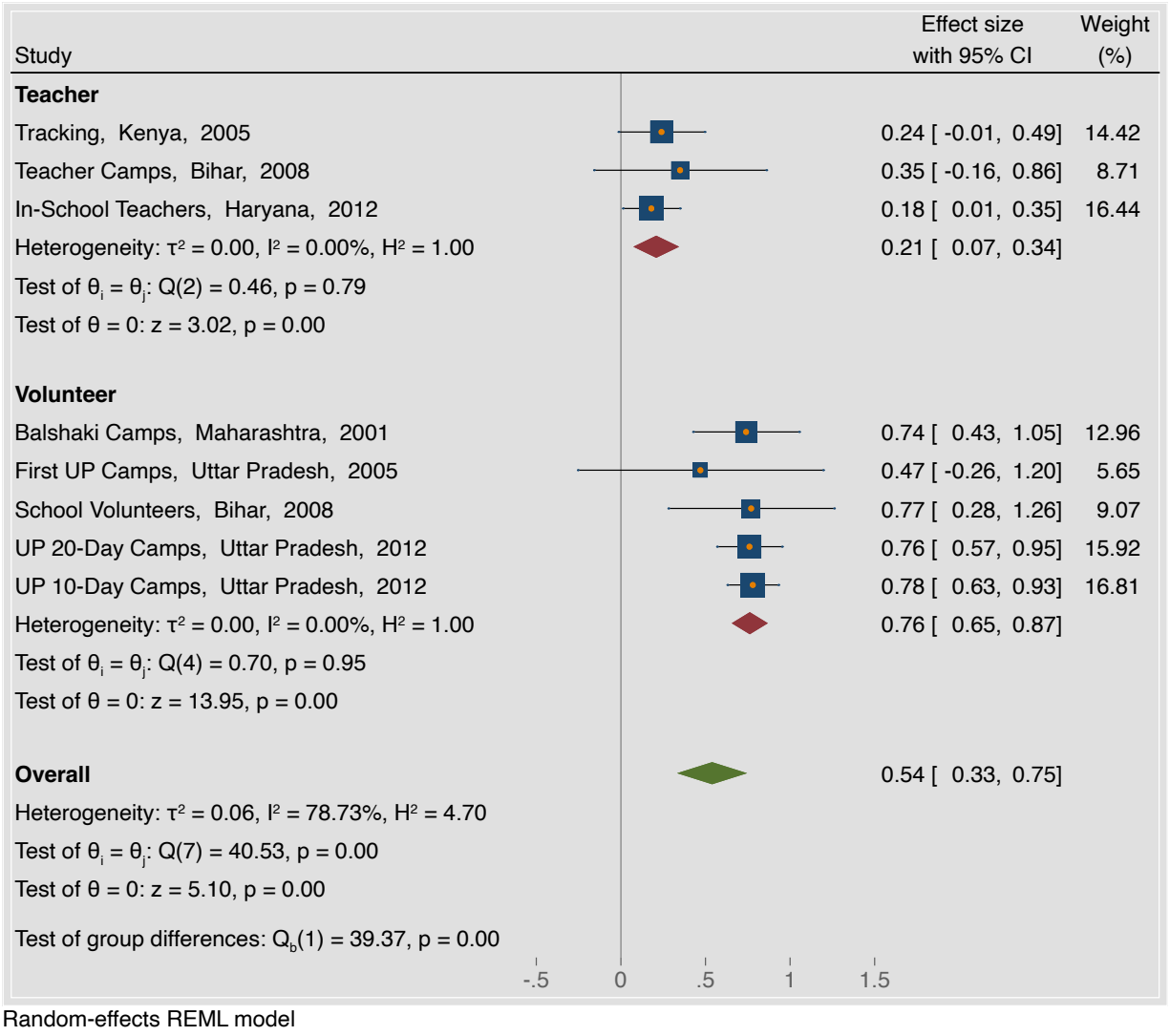

Figure 2: Frequentist Random Effects Meta-analysis of Treatment-on-Treated effects

Second, we aggregate the evidence on the Treatment on Treated (TOT) Effects that we constructed from the microdata. Results are shown in Figure 2. We observe two trends. First both teachers and volunteers are three times as effective conditional on implementation with 0.21 and 0.76 standard deviation average effects, respectively. Both effects are large and precisely estimated. Moreover, we now observe convergence among volunteer effects, with an I-squared of zero. This reveals that much of the heterogeneity in the original volunteer estimates was due to variation in implementation.

These results indicate the initial degree of generalizability of targeted instruction. Effects are large and highly generalizable, conditional on two implementation factors, implementation delivery model (teachers or volunteers) and the degree of implementation (ITT or TOT). The high generalizability after conditioning on these factors leaves little room for a role for other features, such as baseline learning levels, which we later explore formally via metaregression.[17]

[17]It is also worth noting that the patterns in this analysis show few diminishing returns to program scale, suggesting that effects might persist as the program is scaled up: 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. 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.

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. We incorporate wide priors centered at zero to regularize the estimation given the limited number of studies available in such aggregation exercises. 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. In the Appendix we present posterior treatment effects for expected results in future settings, as shown in Figure A3.

Figure 3 displays the results of fitting the basic hierarchical model to the ITT effects of all studies, and Figure 4 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 their heterogeneity across settings. Further confirming this, in an Appendix tables A3 and A4, we report Bayesian pooling factors which are the reciprocal of the I-squared metric. The ITT sees very little pooling, so most of the variation is true treatment effect variation.

However, the Bayesian model pools the TOT estimates to a substantial degree (Appendix tables A3 and A4). Correspondingly, the precision of each study's TOT estimate is enhanced significantly. For example, "First UP camps" effects which is positive but not significant in the no-pooling case is now statistically significant under partial pooling. This is likely due to low implementation in this setting (only 8 percent of students attended sessions) so TOT effects are hard to estimate and inherently noisy without pooling. Pooling studies with high implementation more precisely captures information about the latent effect under full implementation. When average TOT effects are relatively homogeneous across studies, as is true in our case, Bayesian aggregation pools TOT estimates where implementation is low with TOT estimates with high implementation studies, enhancing precision substantially.

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

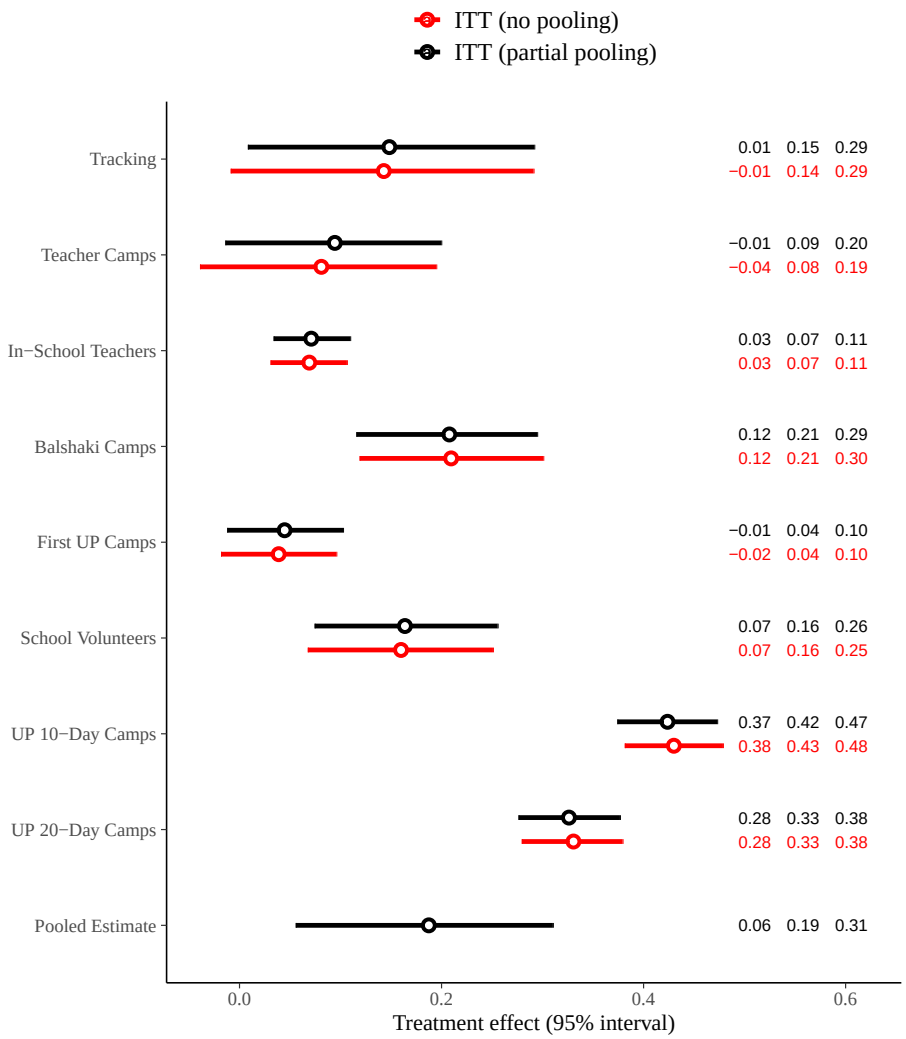

Figure 3: Bayesian Aggregation of all ITT results

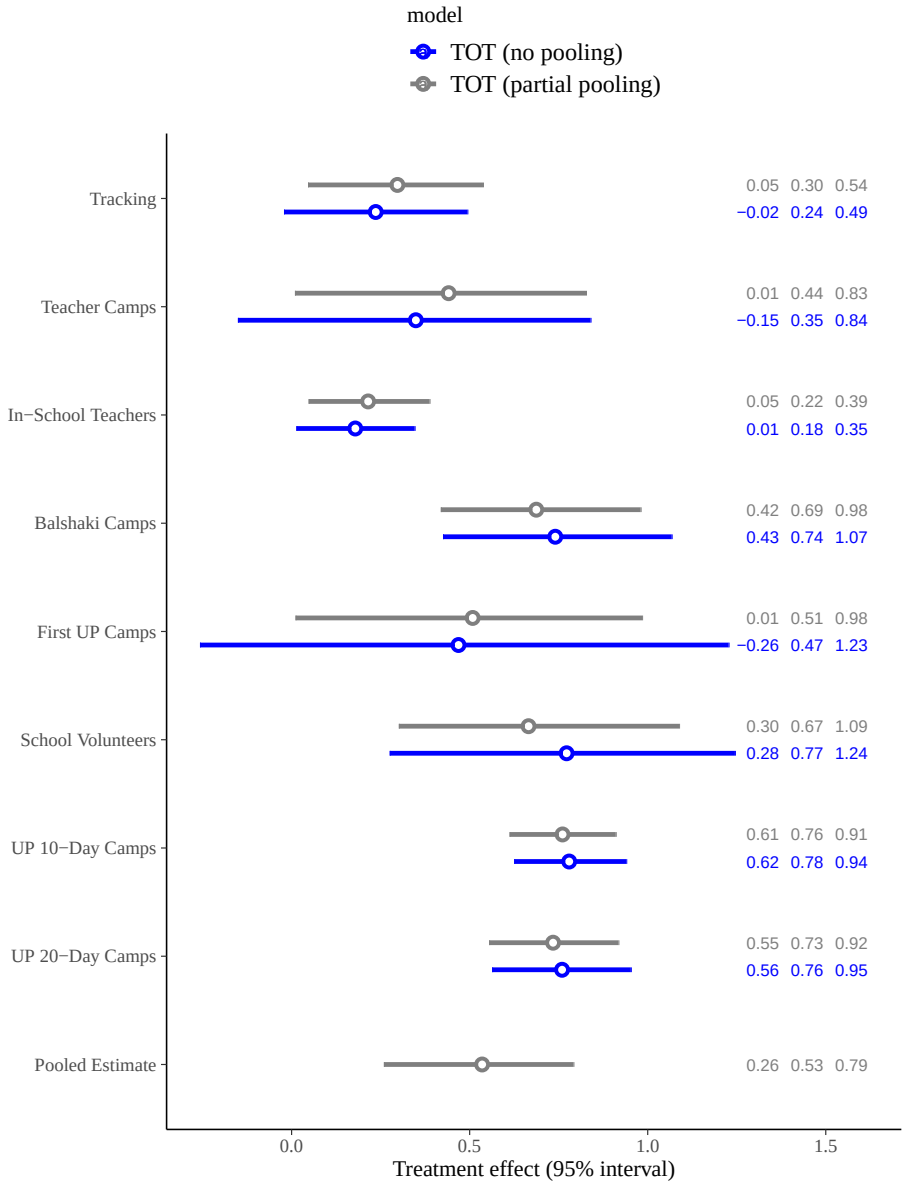

Figure 4: Bayesian Aggregation of all TOT results

We next investigate the role of implementation delivery model (volunteer or teacher) fitting the Rubin (1981) model to each subset – teachers versus volunteers – separately. We present ITT results split by delivery model in Figure 5 and TOT results split by delivery model in Figure 6. The findings show the importance of the delivery model, especially for the TOT results where we see even greater pooling due to the substantial similarity in effects. The visual clustering suggests that when we account for implementation (TOT vs ITT effects) and implementation delivery model (teachers vs. volunteers) the treatment effects are highly generalizable. We formalize this analysis in the following section via metaregression.

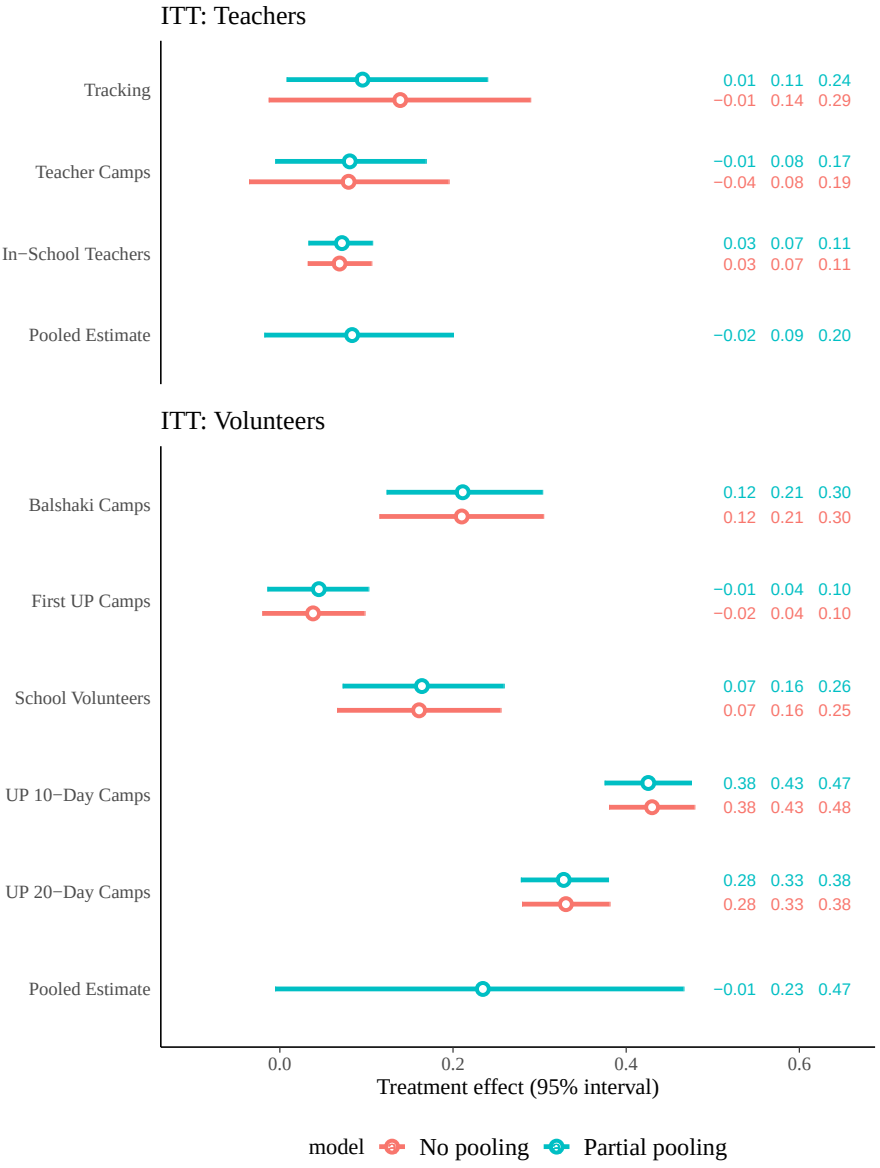

Figure 5: Bayesian aggregation of ITT by implementation delivery model

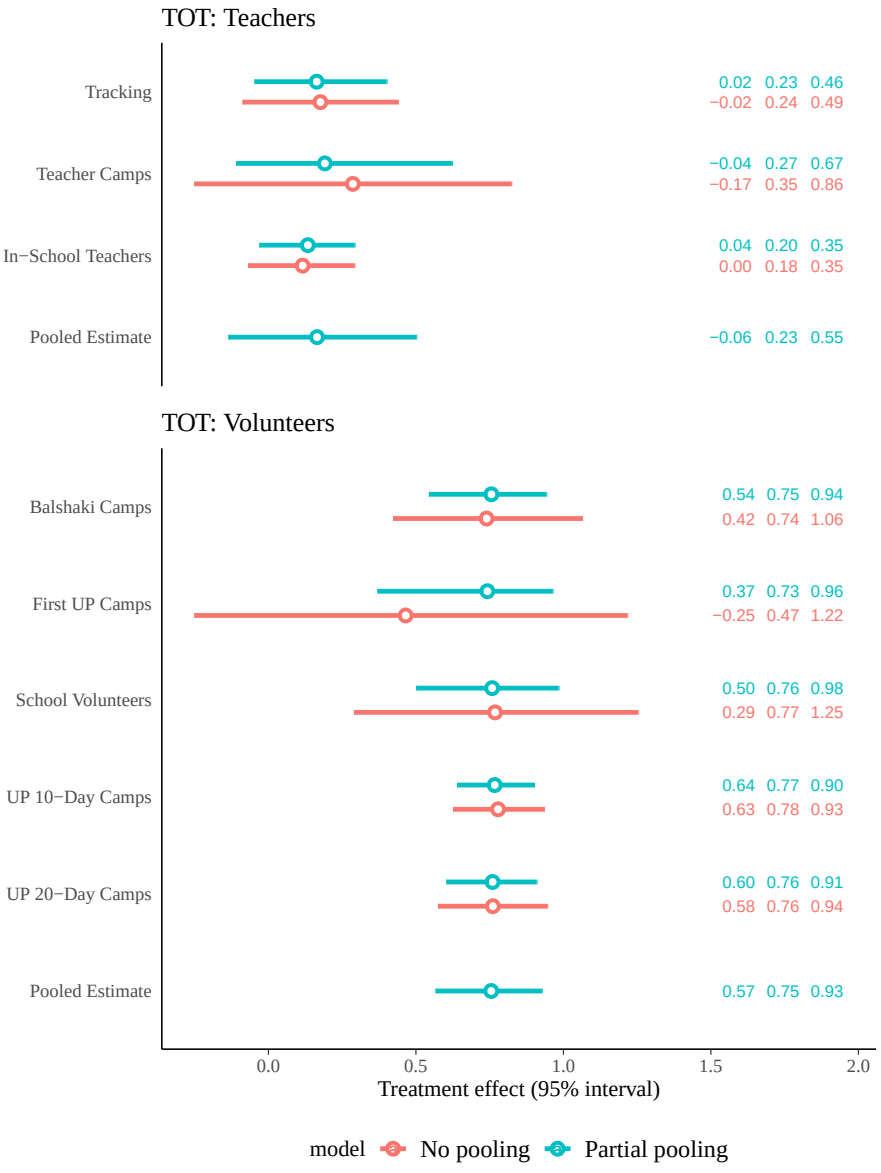

Figure 6: Bayesian aggregation of TOT by implementation delivery model

5.2.2 Bayesian Meta-regression Results

To systematically analyze which factors mediate treatment effects of targeted instruction, we turn to meta-regression. We consider various factors, with a focus on baseline learning levels and implementation delivery model, two of the most probable mediators of large effects. In Figure 7, 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. 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 7 shows, for the ITT, running meta-regression models conditioning on either or both covariates of interest has little impact on the inference. The basic model's 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 explained by these covariates. The TOT model conditioning on implementation delivery model has both the largest average effect of 0.53 standard deviation, as well as the most precise inference (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 single study in the Appendix using a leave-one-out robustness check in Section F.2.

Conditioning on baseline educational performance surprisingly does little to improve precision. One might expect either a negative correlation due to ceiling 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). Yet we see little evidence of any substantial correlation in the data, with this covariate exerting little influence. It is possible that in this data set, all students are so far behind the curriculum that all benefit from targeted instruction, and any additional variation has only a marginal effect. It is also possible that implementation delivery model and the degree of implementation simply play such a large role such that other factors are minor in comparison.

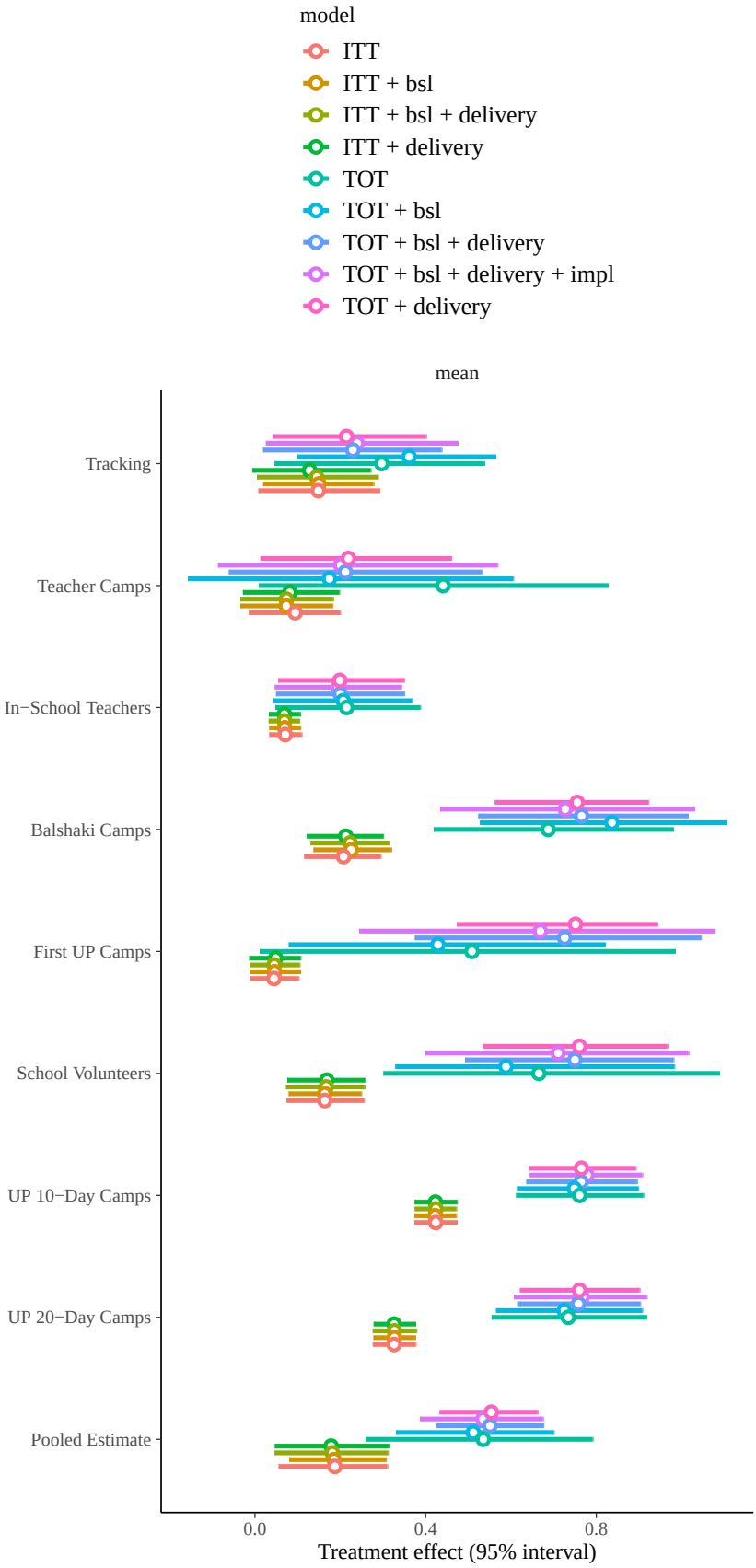

Figure 7: Summary of Bayesian aggregation results across multiple models

6 A Model for Generalizability: The Role of Implementation

Our evidence synthesis reveals the central role of implementation in generalizing effects of targeted instruction. Motivated by these results, we develop a new evidence aggregation model. We first formalize the essential role of implementation information, which we refer to as *m-factors*, in identification of treatment effects and for generalizability. Moreover, the model accounts for uncertainty in implementation.

Our model incorporating implementation information into evidence aggregation introduces a new framework for external validity and generalizability analysis. We show that this model yields more credible results both formally and empirically, enabling substantial generalization of treatment effects across contexts when accounting for implementation. Results are not only more generalizable, they also predict the largest effects in the literature. The results show that targeted instruction can deliver 0.42 SD learning gains on average when taken up, and 0.85 SD gains when implemented with high fidelity, consistent with the upper range of effects found in prior studies.

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 program achievable if it is fully implemented denoted $\theta_j \in \mathbb{R}$. We do not intend this to capture perfect implementation in every detail; rather, we conceive of "full" implementation when the theoretically core components of the program are delivered to intended program recipients on time. We define a notion of implementation that is a proportion rather than strictly binary, since social programs often have multiple core components and may reach some fraction of their program goals even if they do not meet all of them. This is analogous to a notion already embedded in the economics literature: the proportion of recipients who receive the program and defines the wedge between Intention-to-Treat effects and Treatment-on-Treated effects.

We consider two core components of implementation: fidelity to program 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 do not occur then "targeted instruction" did not happen. Hence, we need a notion of implementation that is broader than takeup, and also captures the proportion of instances in which the core components of a program were executed with fidelity.

We define the degree to which a program is implemented in context j as a proportion $m_j \in [0, 1]$, which we call the "implementation factor" or *m-factor*. This factor m_j is 0 when no component of the program is delivered to recipients, and $m = 1$ when the program is delivered as intended to all recipients. When implementation is only partially achieved, we expect to only receive a corresponding fraction of the total potential impact. For example, consider defining implementation as the percentage of time instructors grouped students by ability. Then this percentage is the implementation level m_j, and the latent θ_j is the latent treatment effect of receiving instruction which is targeted the whole time.

Formally, we consider a set of program contexts indexed by $j = 1, 2, 3...J$ and define three relevant objects.

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

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 program is fully implemented.*

Definition 3: Realized Treatment Effect *The Realized Treatment Effect RTE_j is the observed impact of the program in setting j, defined as:*

$$RTE_j \equiv m_j \theta_j$$

Randomized trials recover observed treatment effects, which we define as the *Realized Treatment Effect* RTE_j which equals the Latent Treatment Effect θ_j multiplied by the implementation factor m_j. This structure is analogous to the definition of the Intention to Treat (ITT) effect, which is the Treatment on Treated (TOT) scaled by takeup proportion.[18]

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 program 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 they are just as likely to be influenced by contextual factors as the underlying LTEs are.

6.2 Identification

In this section we establish that Latent Treatment Effects are unidentifiable from Realized Treatment Effects if *m-factors* are unknown. For intuition, consider the case of null effects. Null effects could be due to an ineffective program or an effective program which was never implemented. Without information on implementation, we could misattribute a null to a treatment effect when in fact it is null implementation. More formally, a program that has no treatment effect ($RTE_j = 0$), could be driven by a situation in which $\theta_j = 0$ but, equally possibly, $m_j = 0$. Without explicit information on m_j, a realized effect of zero can not be logically used to infer a null Latent Treatment Effect 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 lay out the formal result to confirm intuition and formalize how relevant parameters can be estimated.

Following Lewbel (2019), we define (point) identification of any parameter as the case in which different parameter values produce observably different distributions of data. We show that the latent effect of a program, θ_j, is not identified from the program treatment effect RTE_j when implementation is not known.

[18] Our notation encodes the assumption that there is a single Latent effect; if this is not the case then our framework can be extended using a set of assumptions similar to those required to extrapolate TOT effects to the broader population, such as the sample receiving the program 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 in many cases 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 program, θ_j, is not identified even if the realized treatment effect RTE_j is observed.*

Proof From definition 3, the realized treatment effect identified in a randomized 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 $\theta'_j m'_j = a\theta_j * \frac{1}{a}m_j = \theta_j m_j$. Thus, by definition 4, 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 program 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 program implementation, if implementation is extremely poor such that $m_j = 0$ then $RTE_j = 0$ for any θ_j; in this case the latent treatment effect of the program is not identified even when implementation is observed. In summary, lack of implementation information or extremely poor implementation prevents the ability to attribute effects to the treatment, and is a major threat to the internal validity of a study.

We now show that a similar identification challenge affects external validity or generalizability when comparing evidence across settings. Recall that we have study settings $j = 1, 2, 3...J$ 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 realized treatment effects $\{RTE_j\}_{j=1}^J$ does not identify the heterogeneity in the set of latent treatment effects $\{\theta_j\}_{j=1}^J$ even when implementation is homogeneous.*

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

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

Since all m_j are not known, this does not identify var(θ_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:

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

The problem arises because anything that scales a random variable's magnitude also scales its variance, and this is true even if the scaling factor is a fixed number with no variation in itself. Naturally, 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 program appear less generalizable 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, or very poor implementation, means that the heterogeneity in the potential effects is not identified.

Since we always observe the realized treatment effects, it may be tempting to wonder how serious this identification problem on latent treatments effects really is. Perhaps it is only realized treatment effects, and not latent treatment effects, that really matter in practice or for policy decisions. Certainly, for those who received the program in the past, the realized treatment effect is all that matters. But for potential future recipients in other contexts, where implementation may be different, the realized effect in previous studies may not be relevant at all. Even for future recipients in the very same context, the realized treatment effect only captures all relevant information if we assume the implementation cannot be influenced or changed. But this is not true: implementation is itself a random variable that researchers and policymakers can affect. We see substantial variation in *m-factors* in our data, ranging from 8 percent to 90 percent, and in Section 7 we show it is possible to improve implementation in the context of a scaling program with large impacts on realized treatment effects. Thus, it is important to quantify *m-factors* and estimate latent treatment effects – which capture the full potential for impact – alongside realized treatment effects. Moreover, these parameters should ideally be jointly studied through a model that can disentangle multiple sources of variation, to which we now turn.

6.3 Incorporating *m-factors* into Bayesian Evidence Aggregation

We now embed the notion of program 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 require extra care to infer. We show in simulations in Appendix Table A1 and A2 that as long as implementation is not exactly zero and we have information about the degree of implementation then it is possible to identify the latent treatment effect 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. We incorporate our model of the realized treatment effect as the product of the latent potential effect θ_j and the associated implementation factor m_j. Because the implementation factor and latent effects are multiplied together, we perform a statistical deconvolution to identify their distributions separately. We observe the estimated realized 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 $\hat{se}_{m_j}$. We now infer the true m_j and θ_j from the data jointly and do so using the model below:

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

To be concrete, consider the definition and measurement of the implementation factor for targeted educational instruction programs. 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 latent treatment effect θ_j would be analogous to "treatment on treated" effects. If we instead define 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 targeted to the right level.

A natural next question is how to define implementation when we have data on multiple aspects of program execution. The answer is to apply the *m-factor* logic recursively: let us say attendance of students in program 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 program can have in half – and if only half the instructors actually deliver targeted instruction, this dilutes the program effect in half again. To perform joint inference on all these factors, the following model may be used:

$$\begin{aligned} \widehat{RTE}_j &\sim N(m1_j m2_j \theta_j, \hat{se}_j^2) \\ \widehat{m1}_j &\sim N(m1_j, \hat{se}_{m1_j}^2) \\ \widehat{m2}_j &\sim N(m2_j, \hat{se}_{m2_j}^2) \\ \theta_j &\sim N(\theta, \sigma_\theta^2). \end{aligned} \tag{6.2}$$

This "2-factor" form of the model allows us to make progress not just on understanding whether implementation matters but which aspects of implementation matter. Conceptually, though not practically, the model above may be expanded ad infinitum. A drawback of our current approach is that we 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 on our small data set. With larger data sets, a richer model with a joint hierarchical structure placed on (m_j, θ_j) could be preferable. Simulations in the Appendix in Table A1 and Table A2, show that 95% posterior interval coverage from our models is generally better than nominal, even when the number of studies is small.

6.4 Implementation Model Results

We estimate Latent Treatment Effects (LTE), first fitting the single-factor implementation model (equation 6.1) to our data. We consider takeup of the program as the level of implementation, as this variable is observed in most studies. Table 2 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 5 posterior quantiles to give the full sense of the distribution, and report the Rhat criterion as a convergence diagnostic. As the results show, the latent treatment effects for all studies are both much larger than the average realized effects and more generalizable, but the difference is much more marked for volunteer studies. The average latent treatment effect for volunteer-delivered programs is 0.49 standard deviations, compared to 0.24 SDs for teacher-delivered programs. Per simulations in the Appendix in Table A1 and Table A2, we use the posterior *median* of the hyperSD as our preferred estimator for this parameter, and we find an approximate hyperSD of 0.23 SD units for each of the delivery models. This implies that latent treatment effects for teachers are likely to be positive in most settings, whereas for volunteers they are always large and positive.

Table 2: Model with Takeup as Implementation factor: Posterior Distribution on Effects

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

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

We note that the inferred latent treatments effects are somewhat smaller than the direct TOT analysis results from previous sections. This is likely because this model accounts for uncertainty on implementation during the aggregation process. This introduces more uncertainty and allows the priors to regularize the estimation towards zero to a somewhat greater extent, as is appropriate in small samples. We also note that the level of takeup in our data ranged from 8 percent to 90 percent; given this substantial range, the linear probability model that underpins the Wald Estimation of the TOT effects is likely to be 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 program fidelity as another important aspect of the implementation of targeted instruction. Although we only have this data for 3 study-arms and we view the results below as suggestive, the single-factor implementation model still performed well in

simulations at $J = 3$ with reasonable coverage of the 95% posterior interval on the 2-factor model (see table A2). We consider it appropriate to proceed with caution.

Table 3 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 2-factor implementation model (equation 6.2) to takeup and fidelity jointly. In both cases, we see large latent treatment effects. The results in Panel A show the median latent treatment effect is now around 0.85 SDs, with a hyperSD around 0.39. The inference in Panel B aligns with Panel A and shows even larger latent treatment effects once fidelity is accounted for. Comparing the results in Panel A considering fidelity of Table 3 to the results considering only takeup (Panel A of Table 2), we find an additional 0.4 SD improvement in the latent treatment effects; this is double what we find when we only consider takeup, revealing the importance of considering multiple types of *m-factors*.

Table 3: Model with Fidelity and Takeup as Implementation Factors: Posterior Distributions of Effects

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

Note: This inference is generated by J =3. Rhat is a diagnostic criterion for MCMC convergence with multiple chains in which a value close to 1 indicates good mixing. We use the posterior median as our preferred point estimate per the simulations in section 6.3. Panel B should be treated as suggestive because model performance measured by RMSE is not reliable for J = 3, although the 95% interval coverage is above nominal.

7 Optimizing Implementation: New Evidence from Botswana

The results of our evidence aggregation establish the importance of implementation in determining program results and generalizability across settings. This offers suggestive evidence that if implementation can be changed in practice, the gains in children's learning may be substantial. We next empirically test whether there are concrete ways to increase take-up and fidelity of targeted instruction in the context of a scaling program. We investigate approaches to increase the fidelity of targeted instruction in the case of Teaching at the Right Level (TaRL) in Botswana, where the government is actively scaling and testing the program in partnership with Youth Impact, one of the largest NGOs in the country. As of 2022, 20 percent of schools in the country had been reached, with all primary schools expected to be reached by 2026.

7.1 Intervention and Study Design

In Botswana, Teaching at the Right Level for numeracy lessons 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 4 below shows the highest operation a child can do at baseline in term 1 of the school year in 2020. The plurality of grade 3-5 students, 30 percent, can do no operations ("beginner" level), 28 percent can do up to addition, 20 percent can do up to subtraction, 15 percent can do up to multiplication and only 6 percent can do up to division. As table 4 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 23 percent of students can recognize up to 4 digits, most children cannot, with 45 percent recognizing only up to 3 digits, and 29 percent of students able to recognize up to 2 digits.

Table 4: 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	
0 digits	0.00
1 digit	0.03
2 digits	0.29
3 digits	0.45
4 digits	0.23

The selected lever to increase fidelity of the intervention was to increase the likelihood that children receive instruction that is optimally targeted to their learning level. To test the viability and benefits of such optimizations of targeting instruction, Youth Impact conducted a randomized controlled trial comparing two options to subgroup students. Youth Impact internally refers to this as "A/B testing", conducting regular randomized optimizations every school term. 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., addition class in one room, mulitiplication 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 recognize 3 digits would be in the same small group as addition-level 1-digit students. The new treatment being trialed randomly ("Option B") involves additionally subgrouping students within an operations-level class according to their digit recognition level. For example, addition-level students who recognize 3 digits would be separated from addition-level students who recognise only 1 digit, with instruction further targeted to digit-recognition level.

The trial took place with over 1,069 students across 52 classes in 4 regions in Botswana, randomized at the class level. While the results of our evidence aggregation predict that improved targeting may improve learning beyond standard implementation, it is not obvious learning will improve ex-ante. First, the relationship observed between implementation and effect size across studies in the evidence aggregation could be driven by omitted variables rather than be causal. For example, certain environments might be 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 certain whether similar level subgroups will improve learning outcomes beyond the standard classroom level operation groupings. It is entirely possible there are diminishing returns to targeting instruction – once instruction is targeted enough, there might be little need to target instruction further, with few gains to improved fidelity above a certain threshold.

7.2 Results

Table 5 reports results of the trial, with the data analyzed using a standard linear regression model estimated via ordinary least squares. Results show that additional sub-learning-level grouping improves number recognition by 0.21 standard deviations on average (column 1), with enhanced precision and an effect of 0.22 standard deviations (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 programs have effect sizes typically around 0.10 standard deviations (Evans and Yuan 2020). 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 randomized trial with those observed in the meta-analysis evidence aggregation. These results reinforce the value of increasing implementation take-up and fidelity, and that implementation is not a black-box; rather improving implementation can be rigorously studied, concrete, tractable, and high-return.

Table 5: Results of a Randomized Increase in Implementation Fidelity

	Outcome: Number Recognition		
	(1)	(2)	(3)
Treatment: Sub-Level grouping	0.205	0.225	0.223
	(0.160)	(0.099)	(0.097)
	[0.205]	[0.027]	[0.026]
Baseline Number Recognition		0.611	0.616
		(0.053)	(0.054)
		[0.000]	[0.000]
Observations	1069	1069	1069
Baseline Level Controls	No	Yes	Yes
Region Fixed Effects	No	No	Yes

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 standard deviations using the control group standard deviation.

The marginal cost of the targeted instruction optimization in this trial is small, estimated at just a few cents. As a result, the cost-effectiveness of optimizing targeted instruction 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

Our analysis demonstrates the importance of quantifying program implementation with as much care as we typically quantify program effects. We find that implementation factors explain most of the variation in the effects of targeted instruction programs across settings. In this case, external validity can be enhanced by simply reporting treatment-on-the-treated effects in addition to the more typically reported intention-to-treat effects. Further insight can be gained by using our new evidence aggregation model that formally accounts for uncertainty in implementation and also incorporates different notions of implementation. We build on insights from our synthesis to guide a new trial optimizing implementation in the context of a scaling program, with substantial increases in treatment effects.

Overall, our results show that research on implementation offers meaningful insights about not just average effects but also, crucially, on the generalizability of effects. Our results also demonstrate that implementation can be changed in practice, identifying concrete mechanisms to achieve the largest frontier effects in the literature. Similar analyses can be conducted across

additional programs in the education sector beyond targeted instruction and might also prove relevant in additional sectors beyond education. Further study of program implementation seems promising, motivating the collection of data on takeup and fidelity at a much more extensive level than is currently practiced.

Our results suggest several avenues for future research on targeted educational instruction. Most importantly, it appears targeted instruction has large and generalizable effects across several low- and middle-income settings, suggesting the approach could make a substantial dent in the learning crisis as it scales. In terms of future research, our results suggest the vanguard of research should focus not on entirely new approaches in education, but rather on optimizing implementation of known approaches, such as targeted instruction, at scale. Implementation should be conceived of broadly, including delivery models, such as teacher or volunteer instruction, and degrees of implementation, such as takeup and fidelity. Developing new theory and practice grounded in richer notions of program implementation may be an important avenue for future work. Another question is why volunteer-led programs appear to be so effective relative to teacher-led programs even, and especially, when accounting for implementation. While teacher delivery is still highly effective with 0.23 standard deviation gains when taken up, volunteers deliver 0.75 standard deviations when taken up. Additional research on how to best ensure teacher take-up and fidelity when adopting targeting instruction as well as research on volunteer-led government models, such as national service programs, to deliver targeted instruction at scale seems promising given such large effects.

An open question is why targeted instruction approaches have been so consistently effective in low- and middle-income (LMIC) contexts yet have more mixed evidence in high-income settings. This could be relevant to the literature on "differentiated instruction" (Tomlinson 2014), a term used to describe approaches similar to targeted instruction in high-income settings and which have found mixed effects. Program implementation seems likely to play a role: a recent systematic review highlighted that in many high-income settings "differentiated instruction has been operationalized in many different ways" (Smale-Jacobse et al. 2019). For example, targeted instruction approaches tested in LMICs typically involve dynamic regrouping of students every few weeks, ensuring instruction is always targeted to children's learning levels. In contrast, in many high-income countries, differentiation involves once-off or more sporadic learning assessment and regrouping of students, likely leading to instruction which is not as well targeted to children's learning level. 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.

9 References

Angrist, Noam, Simeon Djankov, Pinelopi K. Goldberg, and Harry A. Patrinos. "Measuring human capital using global learning data." Nature 592, no. 7854 (2021): 403-408.

Angrist, Noam, Peter Bergman, and Moitshepi Matsheng. "Experimental evidence on learning using low-tech when school is out." Nature human behaviour 6, no. 7 (2022): 941-950.

Angrist, Noam, David K. Evans, Deon Filmer, Rachel Glennerster, F. Halsey Rogers, and Shwetlena Sabarwal. "How to improve education outcomes most efficiently? A Comparison of 150 interventions using the new Learning-Adjusted Years of Schooling metric." (2020). The World Bank.

Andrews, Isaiah, and Maximilian Kasy. "Identification of and correction for publication bias." American Economic Review 109, no. 8 (2019): 2766-94.

Andrews, Isaiah, and Emily Oster. 2019. "A simple approximation for evaluating external validity bias." Economics Letters 178: 58-62.

Bandiera, Oriana, Greg Fischer, Andrea Prat, and Erina Ytsma. "Do women respond less to performance pay? Building evidence from multiple experiments." American Economic Review: Insights 3, no. 4 (2021): 435-454.

Bando, Rosangela, Emma Näslund-Hadley, and Paul Gertler. Effect of inquiry and problem based pedagogy on learning: Evidence from 10 field experiments in four countries. No. w26280. National Bureau of Economic Research, 2019.

Banerjee, Abhijit, Banerji, Rukmini, Berry, James, Duflo, Ether, Kannan, Harini, Mukerji, Shobini, Shot-land, Marc and Walton, Michael. 2017. "From proof of concept to scalable policies: challenges and so-lutions, 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., 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, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying education: Evidence from two randomized experiments in India." The Quarterly Journal of Economics 122, no. 3: 1235-1264.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." Science 348, no. 6236 (2015).

Banerji, Rumini. and Chavan, Madhav. 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, Mark S., Laura Damschroder, Hildi Hagedorn, Jeffrey Smith, and Amy M. Kilbourne. "An introduction to implementation science for the non-specialist." BMC psychology 3,

no. 1 (2015): 1-12.

Beuermann, Diether W., Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu Cruz-Aguayo. "One laptop per child at home: Short-term impacts from a randomized experiment in Peru." American Economic Journal: Applied Economics 7, no. 2 (2015): 53-80.

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, Flavio, and James Heckman. "The technology of skill formation." American Economic Review 97, no. 2 (2007): 31-47.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials." Social Science & Medicine 210: 2-21.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 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, Esther. "The economist as plumber." American Economic Review 107, no. 5 (2017): 1-26.

Duflo, Annie, Jessica Kiessel, and Adrienne Lucas. Experimental Evidence on Alternative Policies to Increase Learning at Scale. No. w27298. National Bureau of Economic Research, 2020.

Evans, David K., and Anna Popova. "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 (2016): 242-270.

Evans, David K., and Fei Yuan. "How big are effect sizes in international education studies?." Educational Evaluation and Policy Analysis (2020): 01623737221079646.

Ganimian, Alejandro J., and Richard J. Murnane. "Improving education in developing countries: Lessons from rigorous impact evaluations." Review of Educational Research 86 (2016): 719-755.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. "Many children left behind? Textbooks and test scores in Kenya." American Economic Journal: Applied Economics 1, no. 1 (2009): 112-135.

Global Education Evidence Advisory Panel. 2020. Cost-Effective Approaches to Improve Global Learn-ing: 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, Michael. "Generalizing the Results from Social Experiments: Theory and Evidence from India." (2023). Working manuscript, Pennsylvania State University.

Gelman, A., John B. Carlin, Hal S. Stern Donald B. Rubin. 2004. "Bayesian Data Analysis: Second Edition", Taylor Francis

Gelman, A, Jennifer Hill. 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

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

Higgins JPT, Green S (editors). 2011. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.handbook.cochrane.org.

Imbens, Guido W., and Joshua D. Angrist. "Identification and estimation of local average treatment effects." Econometrica: journal of the Econometric Society (1994): 467-475.

Innovations for Poverty Action. 2018. "Evaluating the Teacher Community Assistant Initiative." Accessed July 19, 2018. https://www.poverty-action.org/study/evaluating-teacher-community-assistant-initiative-ghana

J-PAL. 2013. "Improving learning by increasing motivation, targeting instruction, and addressing school governance." J-PAL Policy Insights.

Kraft, Matthew A. "Interpreting effect sizes of education interventions." Educational Researcher 49, no. 4 (2020): 241-253.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. "The challenge of education and learning in the developing world." Science 340, no. 6130 (2013): 297-300.

Lewbel, Arthur. "The identification zoo: Meanings of identification in econometrics." Journal of Economic Literature 57, no. 4 (2019): 835-903.

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

Meager, Rachael. 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, Rachael. "Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature." American Economic Review 112, no. 6 (2022): 1818-47.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting education? Experimental evidence on technology-aided instruction in India." American Economic Review 109, no. 4: 1426-1460.

Piper, Benjamin, Stephanie Simmons Zuilkowski, and Abel Mugenda. "Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative." International Journal of Educational Development 37 (2014): 11-21.

Pritchett, Lant. The rebirth of education: Schooling ain't learning. CGD Books, 2013.

Pritchett, Lant, and Justin Sandefur. "Learning from experiments when context matters." American Economic Review 105, no. 5 (2015): 471-75.

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

Smale-Jacobse, Annemieke E., Anna Meijer, Michelle Helms-Lorenz, and Ridwan Maulana. "Differentiated instruction in secondary education: A systematic review of research evi-

dence." Frontiers in psychology 10 (2019): 2366.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt, and Emmanuel Jimenez. "The impact of education programmes on learn-ing and school participation in low-and middle-income countries." (2016).

Tomlinson, Carol Ann. The differentiated classroom: Responding to the needs of all learners. 2014.

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

Vespoor, Adriaan. "Pathways to Change: Improving the Quality of Education in Developing Countries". World Bank Discussion Papers 53.

Vigneri, Marcella, Edoardo Masset, Mike Clarke, Josephine Exley, Peter Tugwell, Vivian Welch, Howard White. 2018. "Economics and Epidemiology: Two Sides of the Same Coin or Different Currencies for Evaluating Impact?". CEDIL Inception Paper 10.

Vivalt, Eva. "How much can we generalize from impact evaluations?." Journal of the European Economic Association 18, no. 6 (2020): 3045-3089.

World Bank. 2018. World Development Report 2018: Learning to Realize Education's Promise. Washing-ton, DC.

A Appendix Figures

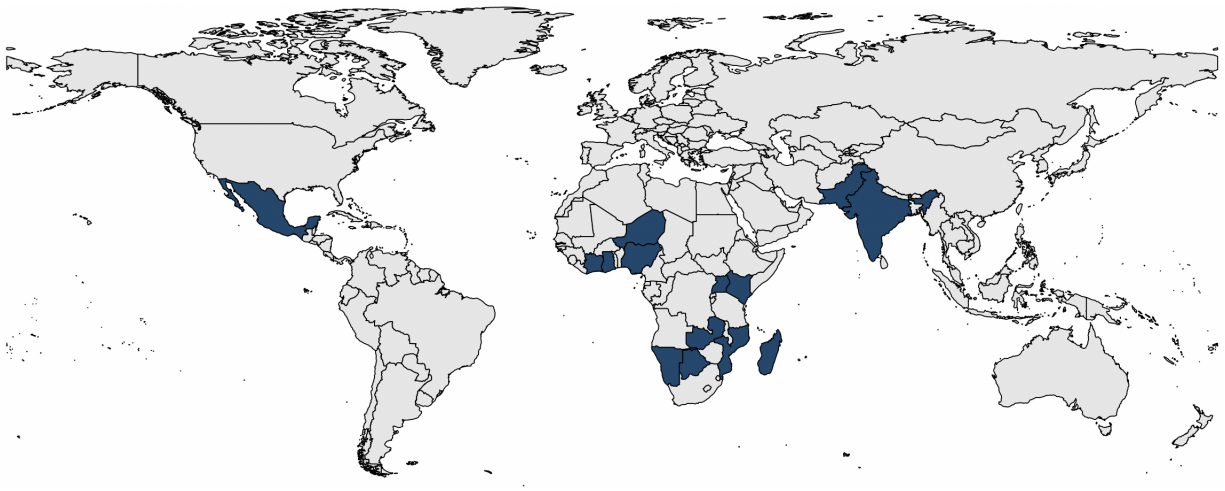

Figure A1: Active targeted instruction scale-up efforts are ongoing in Botswana, Cote D'Ivoire, Ghana, India, Kenya, Madagascar, Mexico, Mozambique, Niger, Nigeria, Pakistan, Uganda, and Zambia. For more information see https://www.teachingattherightlevel.org/)

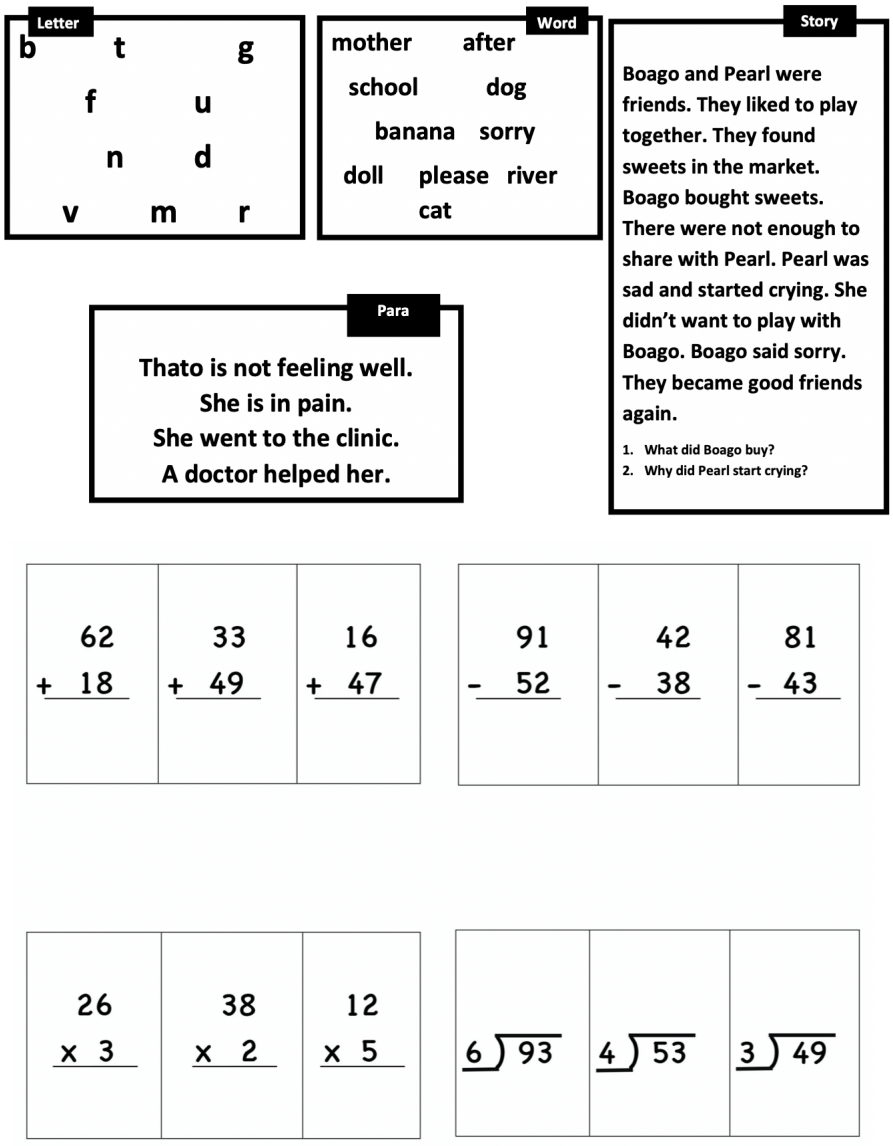

Figure A2: ASER assessment examples used across 14 countries

B 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 A3 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.

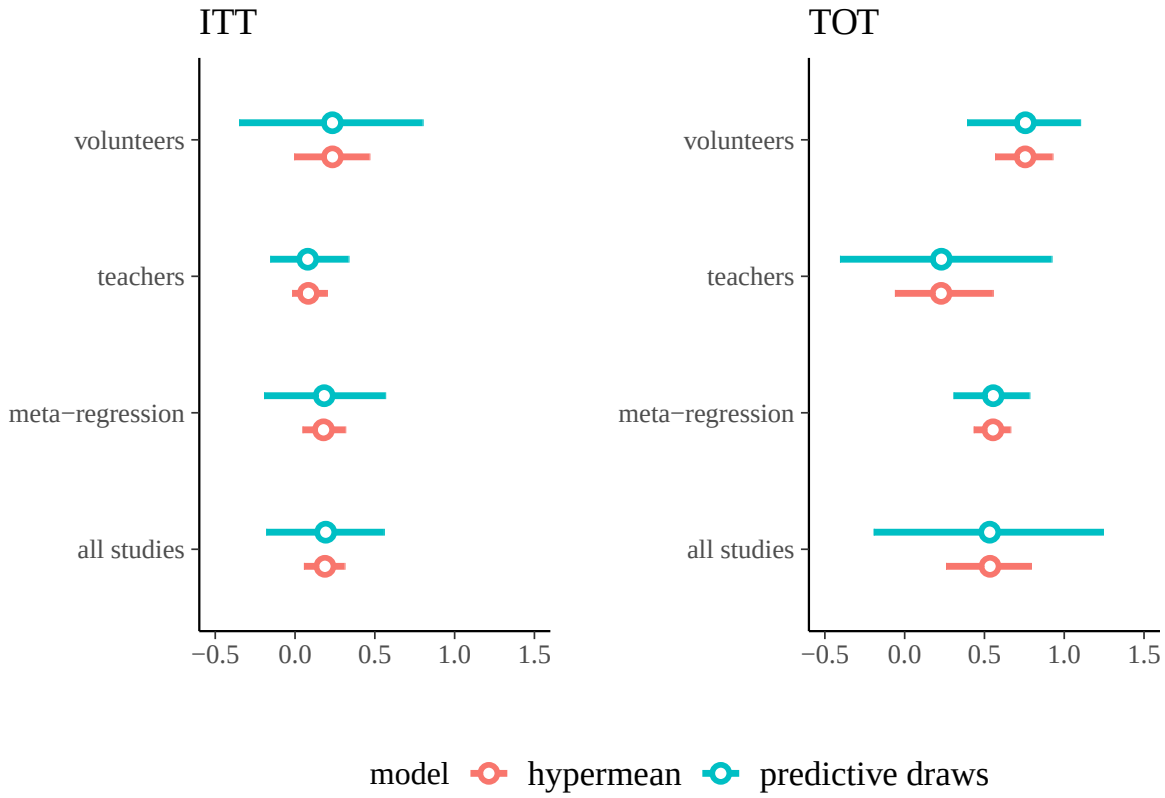

Figure A3: Posterior Predictive Distributions of Future Effects

To interpret figure A3, 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 attempting 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 differential delivery mechanism 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 volunteer TOT effects being both larger on average and, crucially, more homogeneous across studies. In short, our analysis finds strong evidence that volunteer-lead targeted instruction interventions have a generalizably large and positive impact. By contrast, though the average TOT and ITTs effect for all targeted instruction programs 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 A3 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.

B.1 Full Posterior Predicted Distribution Graphics

Figure A4 shows the basic model in red. Meta-regression models conditioning on baseline are shown in green, on implementation delivery model in purple, and on both in blue. As the figure shows, the predicted effect is virtually identical in each case. We examine the 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). This result confirms that the remaining heterogeneity in effects is largely predicted by two implementation factors: implementation degree (e.g. TOT) and implementation delivery model.

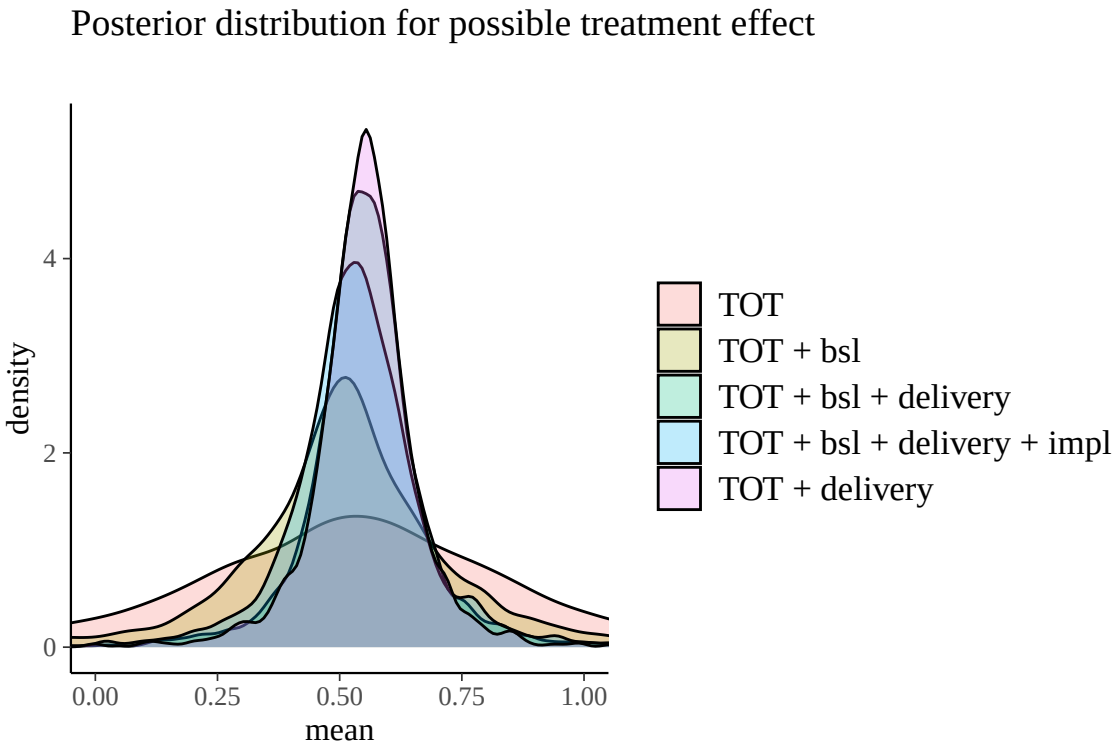

Figure A4: Full posteriors predictive distributions under different models

C Results with Theory-Informed Priors

In this section, we combine the findings of the Bayesian statistical analysis in the previous sections with qualitative expert insight and economic theory. This approach to understanding generalizability bridges economics and epidemiological practice in a manner consistent with the advice for researchers in Deaton and Cartwright (2018) and Vigneri et al. (2018). First, we discuss 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 formalizing 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 grade level. In this environment the curriculum is mismatched to students' zone of proximal development. Moreover, targeted instruction works by creating homogeneous groups which enables more efficient instruction by minimizing the likelihood of mismatch in a given group of receivers of information. This approach is most needed in schooling systems which in the status quo student learning is far behind 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 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 increase the returns of the other; this is behind much of the "big push" development literature that underpins many highly influential development programs including the Millennium Villages Project and the BRAC Graduation program, sometimes called "Targeting the Ultra Poor" (Banerjee at al. 2015). This intuition applies to the targeted instruction intervention: for example, once a child can recognize 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 the students who start at higher baseline levels will progress faster.

Since these theories have qualitative predictions that go in opposite directions, the overall implication for the quantitative estimation model is that one should regularize the correlation between control level (or baseline level) and treatment effects across settings. The translation of 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 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 Teaching at the Right Level 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 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 higher likelihood on moderately

sized correlations of either sign than on extreme correlations of either sign; this offers a smooth, classical regularization in the style of the Ridge penalty.

Literature-informed Priors We now use the theory 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 data set, we limit ourselves to literature published before 1995 (before Pratham was even founded). At that time, the state-of-the-art understanding of experts in primary education interventions in developing countries was overall quite pessimistic about the potential for any single intervention to improve outcomes (Lockheed and Vespoor, 1991). An exception to this general pessimism was surrounding 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, but 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 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. Using a range of strong priors allows us to understand exactly how strong the patterns in this papers' data and statistical analysis are, and also captures insights from a broader set of information beyond the present studies.

Results with theory- and expertise-informed priors Figure A5 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 A6 shows the same analysis for the TOT effects. All priors we consider are Gaussian as discussed in section 4 and centred at a zero effect, yet their strength varies substantially by varying the standard deviation 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 earlier default priors uses, and which are commonly used in the literature, where the prior variance is typically 7.8.

Figures A5 and A6 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 less able to overcome pessimistic priors. Yet all but the most pessmistic prior reports an almost certain positive TOT effect on average, and even with the most pessimistic priors, the Bayesian models report more than 75% chance of a positive TOT effect of targeted instruction. In Section F.1, we also show results using a model that implements regularization

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 evidence of large and positive effects in some studies.

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 there was a pessimism in 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 most promising (Global Education Evidence Advisory Panel 2020). This shift in opinion had to overcome strong priors, and our analysis shows that evidence on approaches such as targeted instruction is indeed strong enough to do so.

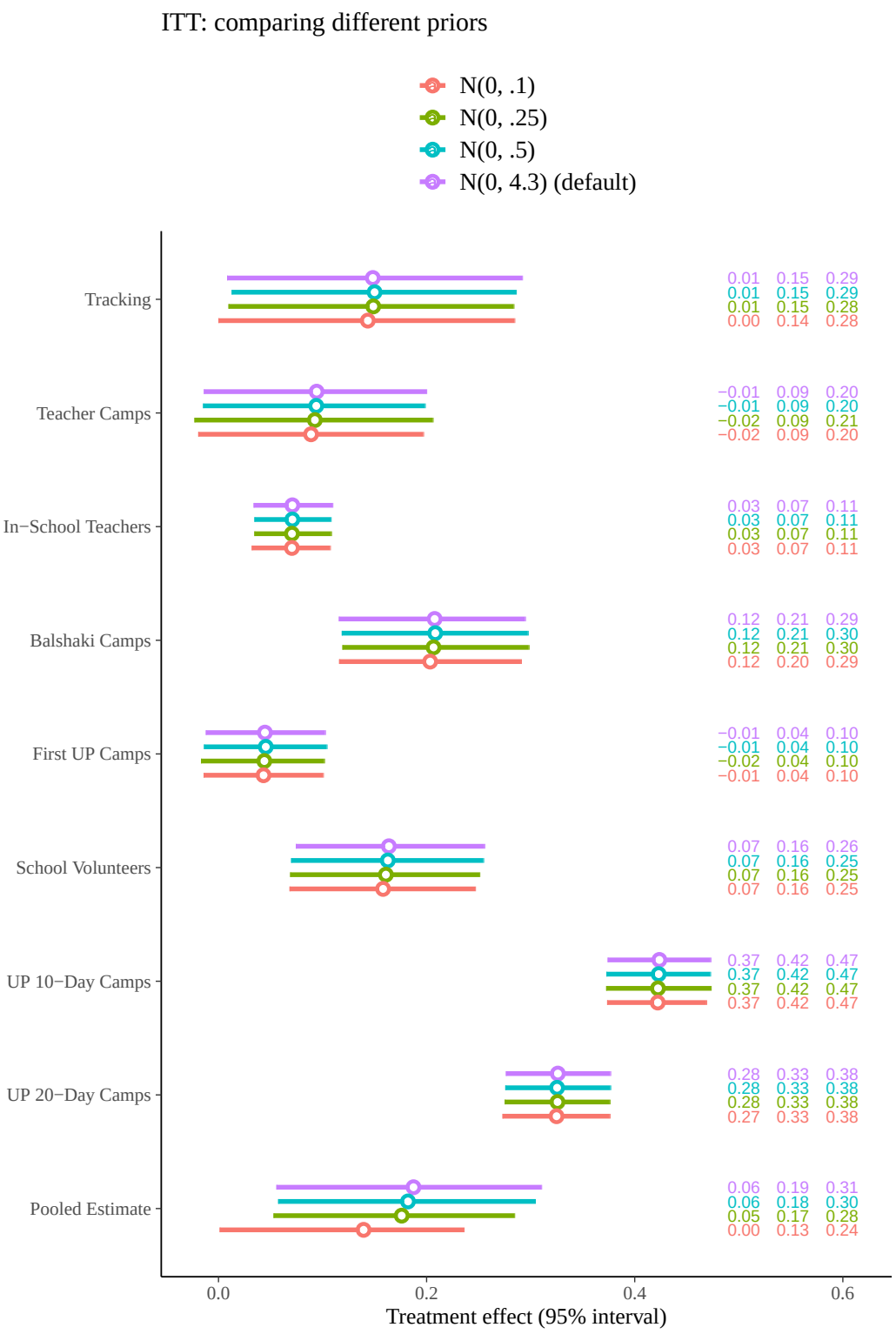

Figure A5: ITT results under different theory-driven priors

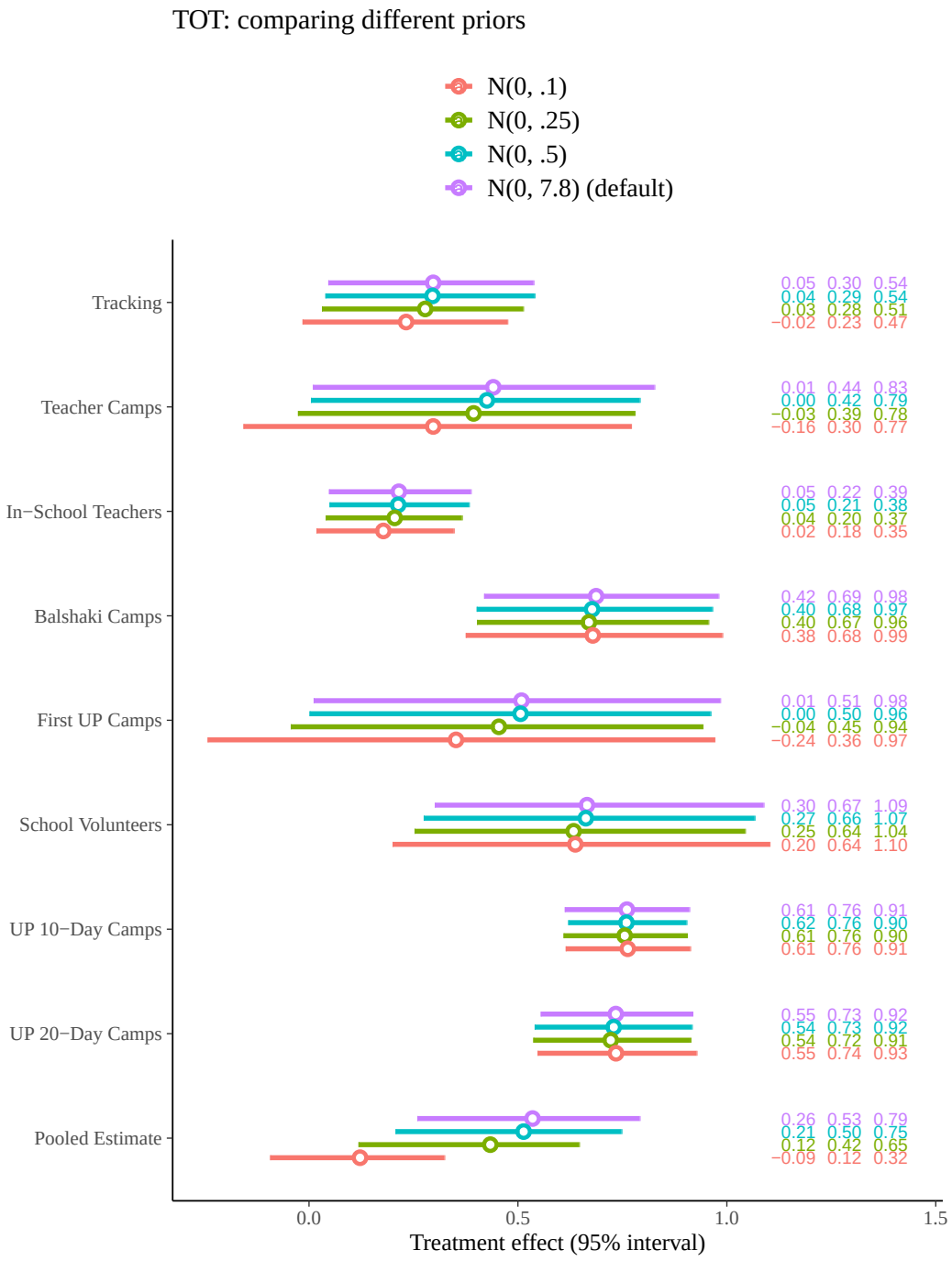

Figure A6: TOT results under different theory-driven priors

D Bayesian Model Estimation Performance

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 data sets 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 realized treatment effects 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 data set and we draw their standard errors from a uniform distribution on [0.005, 0.05] because this is roughly their magnitude in our data set. In each case we record the root mean squared error 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 A1, and in Table A2 for the 2-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 2-factor model, the 50% credible interval's coverage is degraded for the HyperSD when $J < 15$. The results ahow that using the posterior median offers large root mean squared error (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 A1: 1-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 2-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 takeup, using the single implementation factor model is preferable to first computing the treatment on treated 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

Table A2: 2-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

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 takeup in our data ranges from 0.08 to 0.90, and the LATE is underpinned by a linear probability model which is unlikely to perform well over extreme values.

E Additional Bayesian Models and Results

E.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 inference. We specify a joint hierarchy on the control group means and treatment effects in each TaRL study, as follows:

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

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 the program 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 A7. They broadly confirm the Rubin (1981) results shown for comparison.

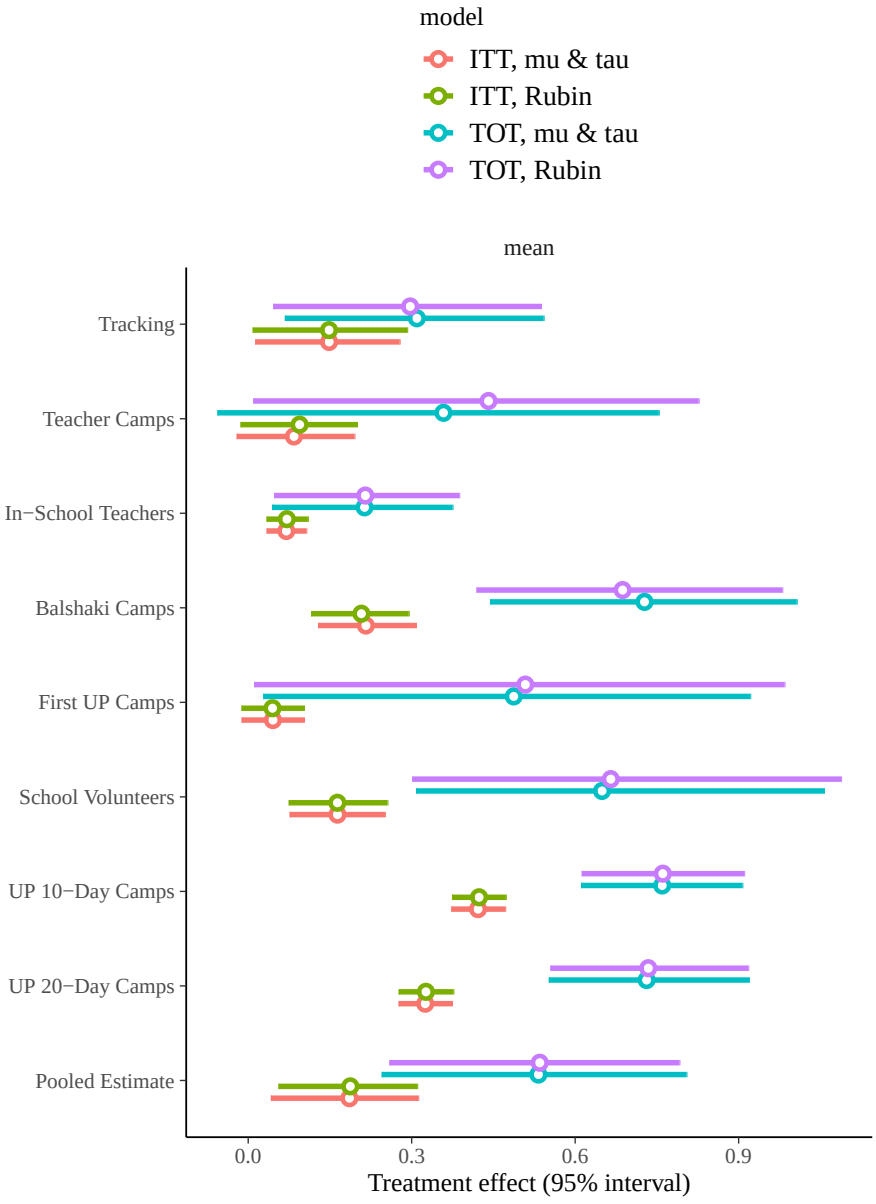

Figure A7: Rubin (1981) model vs joint "mu and tau" model

E.2 Bayesian Pooling Factors

We now display the pooling factor for each study in the simple Bayesian hierarchical model. Table A3 and table A4 show the mean estimated pooling factor along with the lower and upper bounds of the 95% posterior interval. The intervals are moderately tight and while both the Intention to Treat results and the Treatment on Treated results are heterogeneous across settings, the ITT is much more heterogeneous (i.e. with smaller pooling factors). The hierarchical models therefore perform much less partial pooling on the ITTs, even accounting for posterior uncertainty about these pooling factors themselves.

Table A3: ITT Effects: Estimated Pooling Factors

	2.5%	mean	97.5%
Balshaki Camps	0.020	0.092	0.202
First UP Camps	0.008	0.040	0.094
Tracking	0.053	0.207	0.405
Teacher Camps	0.032	0.135	0.285
School Volunteers	0.021	0.095	0.209
In-School Teachers	0.003	0.017	0.040
UP 10-Day Camps	0.006	0.028	0.067
UP 20-Day Camps	0.006	0.031	0.072

Table A4: TOT Effects: Estimated Pooling Factors

	2.5%	mean	97.5%
Balshaki Camps	0.054	0.251	0.544
First UP Camps	0.233	0.600	0.865
Tracking	0.036	0.186	0.441
Teacher Camps	0.131	0.445	0.759
School Volunteers	0.122	0.428	0.745
In-School Teachers	0.016	0.095	0.257
UP 10-Day Camps	0.013	0.078	0.217
UP 20-Day Camps	0.021	0.119	0.310

F Further Robustness

F.1 Prior Robustness results

Figure A8 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 earlier, the TOT effects are estimated with greater uncertainty within each study and therefore 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 evidence of large and positive effects 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.

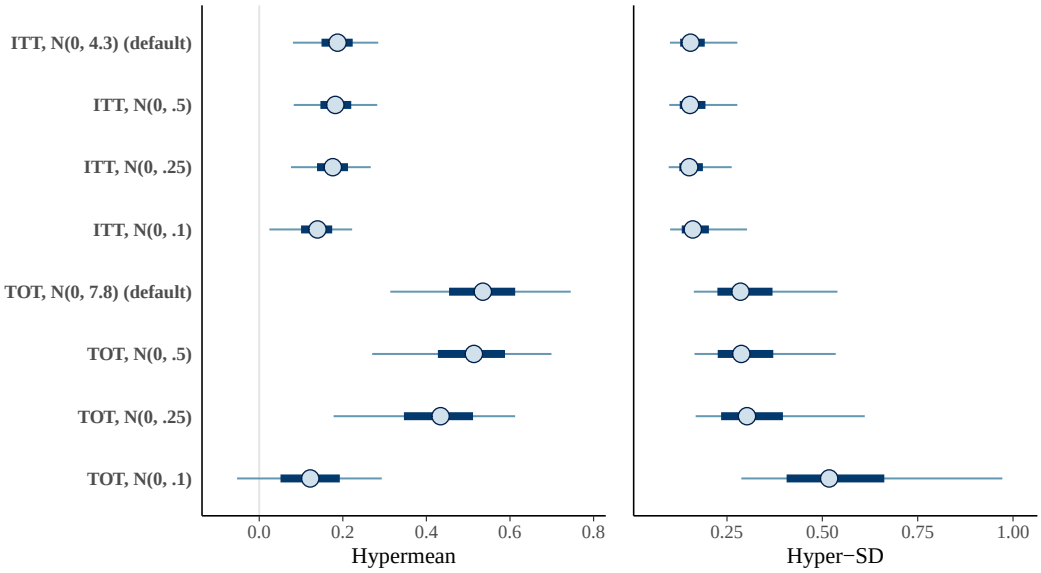

Figure A8: Additional Prior Robustness Checks

We examine the results of imposing a stronger theory-driven regularization of the correlation between baseline educational performance and the treatment effect of targeted instruction towards zero. The competing theoretical mechanisms suggest we should expect a small correlation; this corresponds to expecting or favoring independence or zero-off-diagonal terms in the variance-covariance matrix. This is implemented via the use of an 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, parameterized 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 favors independence, expressed by zero off-diagonal terms. The larger the parameter is, the more strongly it favors independence, and thus, the more strongly it regularizes the correlation

in question.

The graphics in figure A9 below show the results of fitting the joint aggregation model (the "mu and tau" model) with an LKJ prior with concentration 1 (the default used in the previous sections), as well as 3 (moderate regularization) and 6 (strong regularization). In this case the stronger priors have no impact on the posterior hypermeans for TOT or ITT. While this is initially surprising, it reflects a small empirical correlation between the baseline and treatment effect.

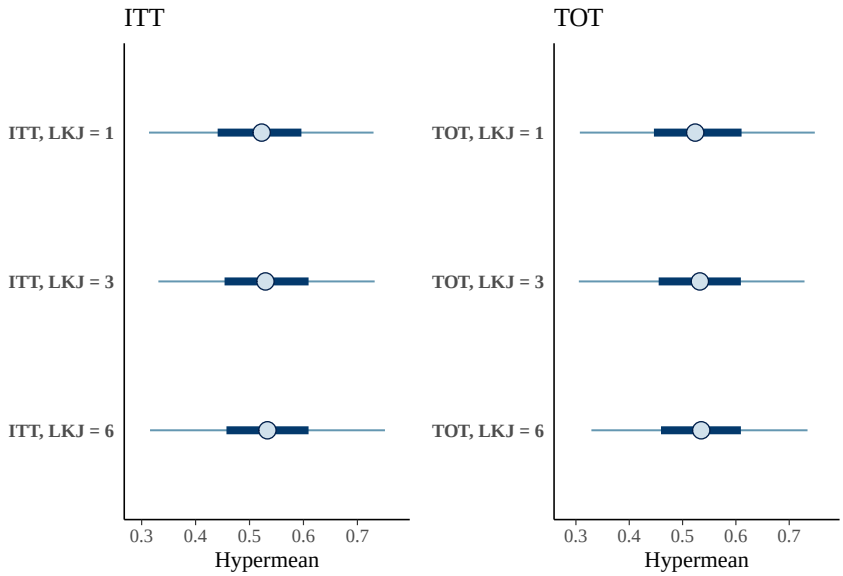

Figure A9: Mu-tau model with different LKJ priors

These results do not necessarily mean baseline levels of learning do not matter for targeted instruction to be effective. Rather it is possible that the set of studies included are all cases with relatively low baseline learning. Thus, if low baseline learning is a critical condition whereby targeted instruction is needed and effective, for all studies this condition might be met, hence positive effects across the board.

F.2 Evidence Aggregation Robustness Checks

In this section 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 implementation delivery model in figure A10, where the study indicated in the row label is the study omitted for that run of the model.

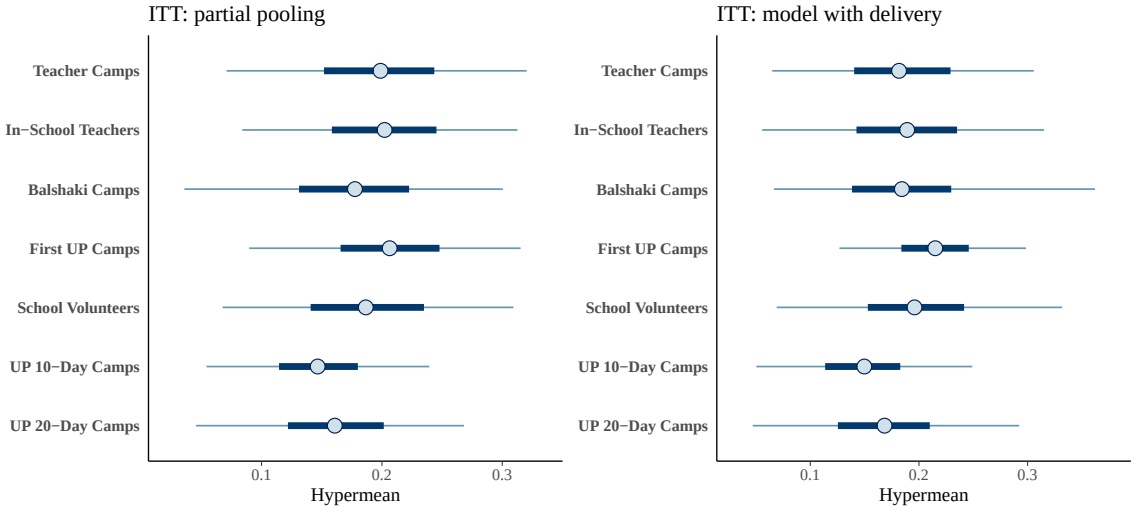

Figure A10: ITT Leave One Out analysis

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 A11 shows the sensitivity results of leaving out studies in turn for the TOT estimates both in the Rubin (1981) partial pooling model and for the model conditioning on implementation delivery.

The graphs clearly show little variation in the posterior distribution of the hypermean for

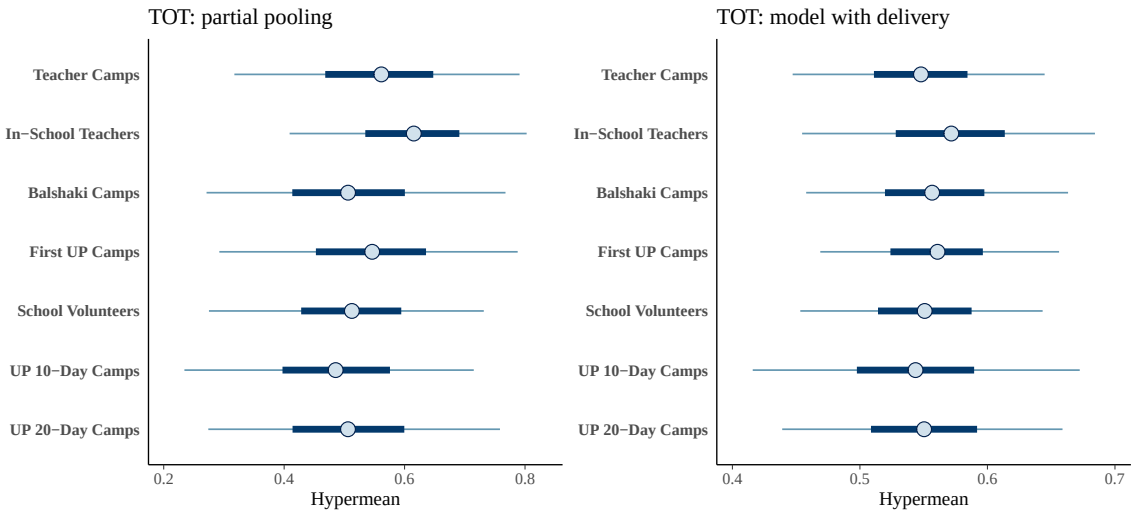

Figure A11: TOT Leave One Out analysis

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 model. This shows the strong robustness of the TOT results to omitting any of the studies, and demonstrates that our insights about the important role of implementation factors are not based on any single study but rather 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 A12 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 cutoff 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 potential for 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.

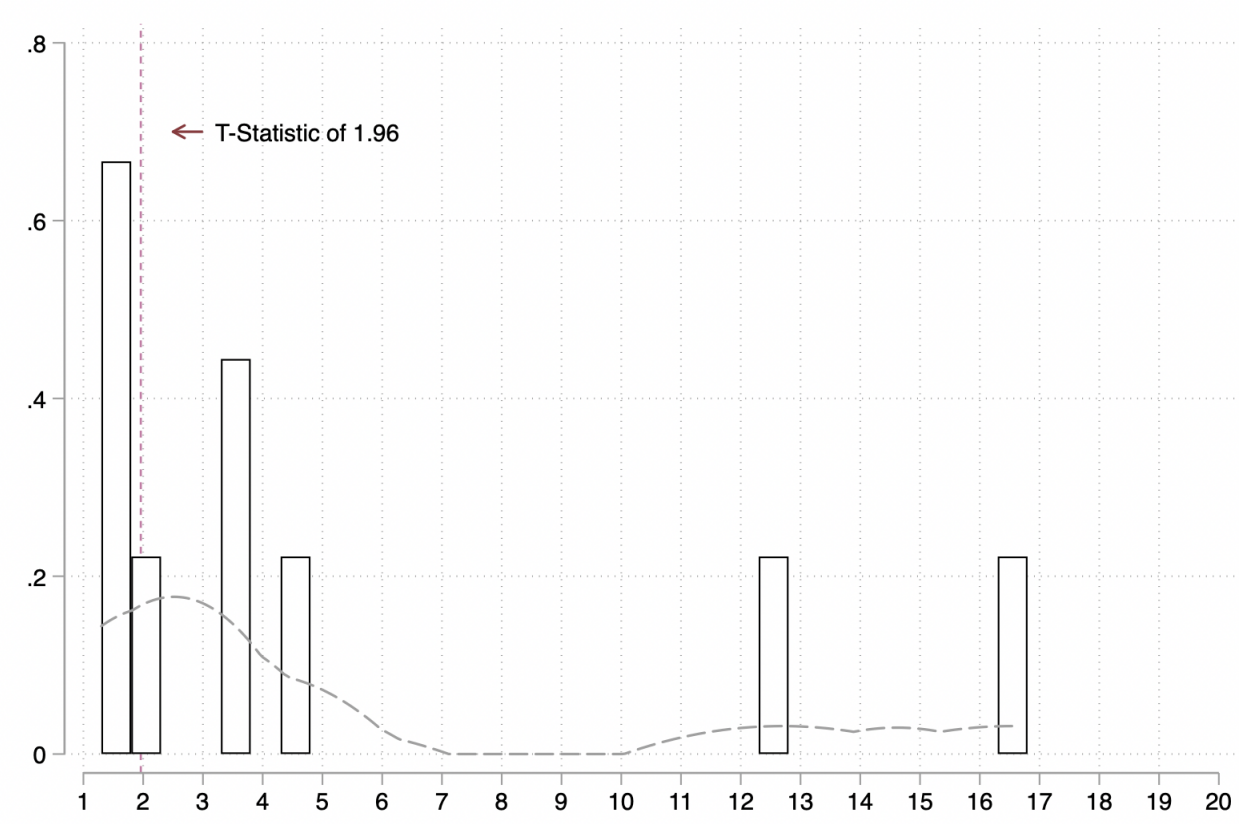

Figure A12: Distribution of t statistics in our sample