Pre-Analysis Plan
Do Students Benefit from Blended Instruction? Experimental Evidence from India*

Andreas de Barros†

December 12, 2018

Abstract

This experimental study investigates the causal effect of a teacher capacity building program that promotes blended instruction, on student learning. It will be implemented in government schools in Haryana, India, in collaboration with a large, local NGO ("Avanti Fellows"). The program's objective is to positively affect the instruction of mathematics and science, in grades nine and ten. The study hypothesizes that student learning improves if teachers are given resources and training, to enrich their instruction with video-based learning materials. Secondly, the study hypothesizes that the intervention's cost-effectiveness outperforms that of an alternative model of teacher capacity building, which does not rely on infrastructure upgrades and uses printed workbooks only.

JEL codes: C93 Field Experiments; I21 Analysis of Education; I22 Educational Finance; I25 Education and Economic Development.

*First and foremost, I gratefully acknowledge the Government of Haryana and staff at Avanti Fellows for their contributions to the "Sankalp" project, and for their outstanding commitment to evidence-based education policy. This research receives financial support from the J-PAL Post-Primary Education Initiative (PPE). The Michael and Susan Dell Foundation supports Avanti Fellows with funding for the "Sankalp" project and its evaluation. I thank Felipe Barrera-Osorio, Alejandro Ganimian, Karthik Muralidharan, and Martin West for continued feedback that informs this study. I also thank Heather Hill, Sharon Kim, Matthew Kraft, Ezequiel Molina (and the World Bank's SABER team), Edward Seidman, and Carmen Strigel, for making instruments available. I thank Brinda Sapra and Anuja Venkatachalam for providing excellent research assistance. The usual disclaimers apply.

†Ph.D. Candidate, Harvard University, Graduate School of Education. E-mail: adebarros@g.harvard.edu.

Contents

1 Introduction

This experimental study investigates the causal effect of a teacher capacity building program that promotes blended instruction, on student learning. It will be implemented in government schools in Haryana, India, in collaboration with a large, local NGO ("Avanti Fellows"). The program's objective is to positively affect the instruction of mathematics and science, in grades nine and ten. The study hypothesizes that student learning improves if teachers are given resources and training, to enrich their instruction with video-based learning materials. Secondly, the study hypothesizes that the intervention's cost-effectiveness outperforms that of an alternative model of teacher capacity building, which does not rely on infrastructure upgrades and uses printed workbooks only.

This research thus seeks to contribute to three main areas. First, the study relates to literature on innovative, "untraditional" teacher training approaches, which—among other program components—stress continued interaction with trainees (Egert et al. 2018; Kraft et al. 2018). Secondly, this research relates to evaluations of education technology in less-developed countries (more specifically, as a means to improve instructional quality through partially scripted lessons) (Gove et al. 2017). Finally, this study also speaks to the question of how promising educational interventions can be effectively administered at scale (Banerjee et al. 2017; Muralidharan and Niehaus 2017; Vivalt 2017).

The goal of this pre-analysis plan is to precisely pre-specify the study, its scope, and its analysis. The document is structured as follows. I begin by describing the program and its components. This is followed by a description of the sample of schools, their randomization into three groups, and calculations of statistical power (i.e., the study's ability to detect the program's effects on student learning outcomes). The next section details data sources and strategies that will be used to measure the program's main outcomes, its intermediary effects, its implementation fidelity, and its costs. Subsequently, I present how the program's causal effects will be assessed econometrically. In doing so, I focus on the intervention's primary outcomes (i.e., effects on student learning); however, analyses of effects on student sub-groups and on program mechanisms are also specified, as secondary work. The pre-analysis plan concludes by presenting its reporting plan, and by providing additional psychometric details in a short, technical Appendix.

2 Experiment

2.1 Intervention

240 Government Senior Secondary Schools (GSSS) across 8 districts of Haryana will be equally randomized across two treatment arms and the control group, with 80 schools assigned to each group. The following provides a more detailed description of the intervention, for the two treatment arms. This section also describes the comparison or "control" group, in which schools continue their operations as usual.[1]

1. **ICT Group**: This group will receive the full intervention. This includes setting up smart classrooms, provision of digital content to supplement teaching instruction, printed workbooks for practice of students, and capacity building of mathematics and science teachers responsible for teaching class ninth and tenth curriculum. Two smart classrooms enabled with ICT (Information and Communications Technology) infrastructure will be set up per school. Procurement, installation and maintenance of projectors, computers and sound systems will be under the purview of HSSPP (Haryana School Shiksha Pariyojna Parishad). Capacity building workshops for teachers will be designed and implemented by Avanti Fellows.[2] Practice workbooks designed by Avanti will be printed and distributed by HSSPP across 160 schools in the two intervention groups.

2. **Workbook Group**: This group's program components are equivalent to those administered in the previous group; however, the group does not receive those particular program components related to ICT (i.e., ICT-related infrastructure upgrades or digital content).

3. **Control Group**: This group continues with "business-as-usual". The schools assigned to the control group will neither receive facilities nor materials. Their teachers will also not undergo the program's teacher training activities.

Either of the two intervention groups allow for a comparison of the program with what is currently considered "business-as-usual", for Haryana government schools (by contrasting groups 1 and 2 with group 3, above). In addition, the study will provide information on the program's cost-effectiveness through a comparison of the "full" (costlier and more difficult to

[1] Table 1 in Appendix B provides a detailed breakdown of responsibilities, for the two implementation partners, along with the program's components and intervention groups. To facilitate the program's replicability and scalability, the study will provide an even more detailed description of the program, following Arancibia et al. (2016).

[2] The initial workshop will be for three days, followed by on-site training for ten days.

implement) intervention with an alternative model that does not rely on ICT. To do so, the study will compare the effects in groups 1 and 2, above. The following sections provide a more detailed description of the randomization strategy, as well as the study's analytic approach.

2.2 Sample, randomization, and power calculations

The study includes all eighth- and ninth-graders in 240 schools, in eight Haryana districts, as selected by the Government and by Avanti Fellows (these students will graduate to grades nine and ten, shortly after). The study will cover two cohorts of students – a first cohort that starts (towards the end of) the school year 2018/19, and a second cohort that starts (towards the end of) the school year 2019/20. In the first cohort, students in grade 8 will be followed for two years. All remaining students will be followed for one year.[3] In the two grades, average enrollment is a combined 122 students, per school (51 in grade 8, 71 in grade 9), for an expected total of 29,280 students in the first cohort, and an additional 12,240 students in the following cohort.[4]

Schools are assigned with an equal split across the three groups (for a total of 80 schools per group). To achieve similar control and treatment groups and to improve statistical power, randomization is stratified. Within districts, schools are sorted into randomization strata of three, grouping schools with similar performance on the Haryana Board Exams together.[5] If the number of schools for any district ranking is not divisible by three, I randomly assign the remaining schools. In this case, I assign these remaining schools globally, i.e. *across* districts (cf. Carril 2016).[6] Finally, I repeat the above randomization strategy ten times, selecting the randomization with the smallest t-statistic, from comparisons across the treatment and

[3] In the second cohort, students in grade 9 are those first-cohort students who graduated from grade 8 and thereafter attend grade 9. At the time of pre-registration, due to funding constraints, there is still no clarity as to whether students of the second cohort can be followed for two years.

[4] Information as per a feasibility audit conducted for the program, in 2018. In grades 9 and 10, the average combined enrollment is 152 students, per school.

[5] To rank schools, I use the average school-level Board Exam score for 2018, overall (covering English, Hindi, Mathematics, Sanskrit, Science, and Social Studies). More precisely, I calculate a school's weighted average, using information on the number of student in each of six performance ranges, on the test.

[6] As more than three schools need to be randomized across districts, I once more group schools with similar Board Exam scores together.

control groups (for a set of select covariates[7] and the average school-level Board Exam score for 2018) (cf. Bruhn and McKenzie 2009).[8]

Conservatively, the study will thus be powered to detect an intent-to-treat effect of 0.158 standard deviations (SD). This number should be considered an upper bound as it focuses on a single cohort of students, only (students in grade 8, at baseline, to be followed over two years). Figure 1 represents this calculation visually, plotting Minimal Detectable Effect Sizes (MDEs) against the number of schools in the study.[9] As shown in the figure, there are substantial benefits from using 240 schools, as opposed to a smaller sample (such as 180 schools). To put the minimal detectable effect into perspective, in my work with Ganimian and Muralidharan (2017) in Rajasthan, we find a 0.47 and 0.33 SD difference between sixth- and seventh-graders, and seventh- and eighth-graders, respectively (in math). Among students in the set of 59 pilot schools, in Rajasthan, I find a 0.26 SD difference between ninth- and tenth-graders (in math).

3 Measurement and data

The following describes outcome variables and their measurement. Data integrity and data security will be guaranteed through standard quality checks, including (out of sample) piloting, (back-)translation of instruments between English and Hindi, spot-checks and accompaniments, high-frequency checks, and the use of end-to-end encryption (for digital data-collection) (cf. Glennerster 2017).

Student learning. The study's main program outcome of interest is student learning, in math and science. Students in grades nine and ten will be assessed at baseline (in December) and approximately one year later, through an end-of-year test (in January).[10] Assessments

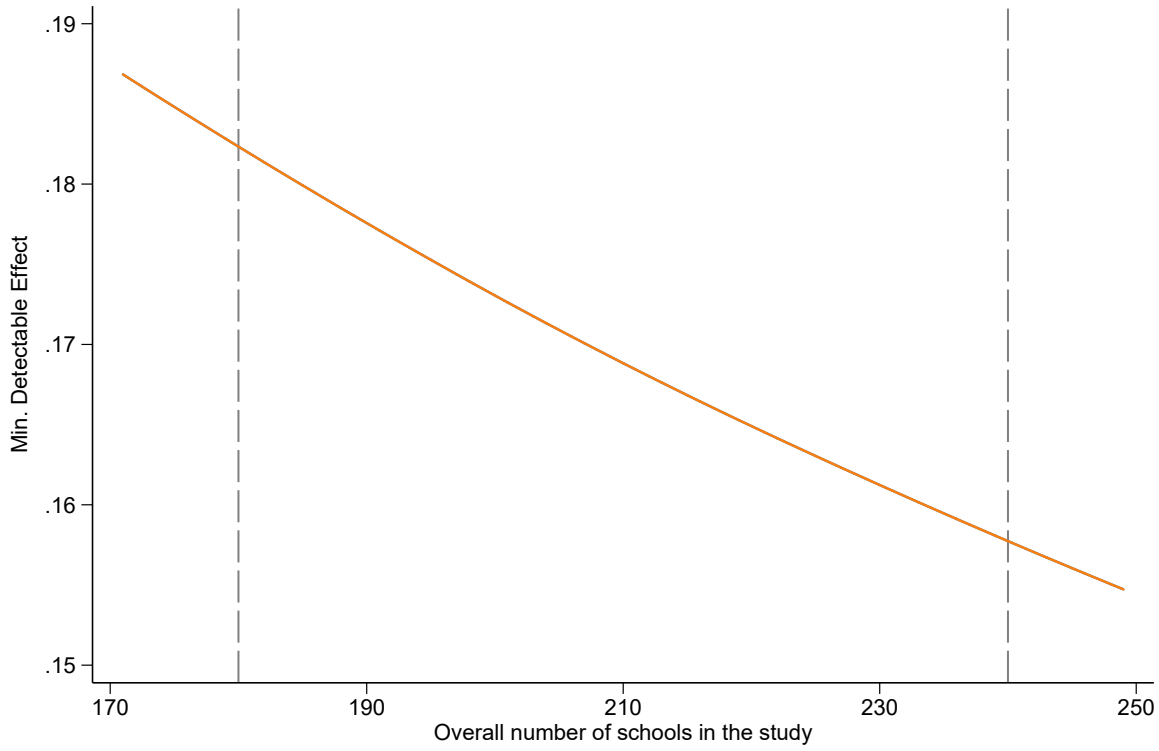
[7]Using Lasso, I select a set of covariates from India's District Information System for Education (DISE), which are predictive of school-level average pass rates for standardized exams (in grades 7 and 8). This procedure selects the following covariates: Number of students in grade 7 and 8; percentage female (students); percentage minority (students); percentage "Other Backward Class" (students); number of teachers; percentage female (teachers); percentage graduates (teachers); years in service (school); co-ed (school); school requires minor repairs; wall missing or damaged; ratio computers/students.

[8]Comparisons across groups account for stratification fixed effects. I follow Athey and Imbens (2017) and Banerjee et al. (2017), who suggest that the number of re-randomizations should be limited and pre-specified. I am not aware of an optimal number, but consider 10 re-randomizations as conservative.

[9]I compute the MDE in standardized test scores as a function of the number of schools, for analyses at a 5% significance level, with 80% power. Based on assessments conducted with 59 pilot schools in the same state, in 2017/18, I assume that covariates will account for 36% of the endline variation. I also account for a (worst case) scenario, in which attrition will be 35%. Based on the same set of assessments, in 59 pilot schools, I assume an intra-school correlation of scores of 0.17 (after accounting for students' baseline scores). Given school-level assignment, I assume absence of spill-over effects.

[10]Students who take their baseline assessment in grade 9 will take two follow-up tests: one at the end of grade 9, and another at the end of grade 10. Students who take their baseline assessment in grade 10 will only take one follow-up test, at the end of grade 10.

Figure 1: Power calculations



Note: This figure plots Minimal Detectable Effect Sizes against the number of schools in the study. The vertical lines indicate 180 and 240 schools, respectively. MDEs reflect main intent-to-treat effects. 1/3 of schools in the ICT Group; 1/3 in the Workbook Group; 1/3 doing business-as-usual. Calculations assume power=0.8, alpha=0.05, spill-over=0, attrition=0.35, R-squared=0.36, ICC=0.17, 51 8th-graders/school. These assumptions reflect Ganimian et al. (2017), and additional pilot results from 59 schools, in Rajasthan.

will be administered as paper-based tests (one per grade and subject), under the same, strict governmental oversight as other central exams (with additional monitoring from the research team). Test items are tightly mapped to the official "CBSE/NCERT" school curriculum, but also include items from up to two years below grade-level. Items have been administered in similar contexts previously, in large-scale assessments. From these previous administrations, item response theory (IRT)-based item characteristics are used to maximize each assessments' test information.[11] Estimates of student ability will be calculated using a standard, three-parameter logistic (3PL) IRT model, with a single guessing parameter (Birnbaum 1968; Samejima 1973).[12] In doing so, anchor items across grades and test-occasions will allow for the (concurrent) linking of estimates onto one common, continuous ability scale per subject (Kolen and Brennan 2004; Stocking and Lord 1983).

[11] See Jacob and Rothstein (2016) for an accessible introduction to item response theory, in the economics literature.

[12] In case a 3PL model does not converge, a 2PL model will be used instead.

Teaching behaviors and quality of instruction. The program's effects on teaching behaviors and on instructional quality will be assessed through two instruments: Classroom observations and student reports. First, the program includes bi-weekly school visits (conducted by government "block officers") and monthly classroom observations (through Avanti Fellow's field staff). All visits and observations will be conducted in intervention and control schools. For this purpose, the study developed an instrument to measure the program's effect on the quality of instruction a student receives.[13] Secondly, during school visits, a subset of students will be surveyed on common classroom behaviors. The study explicitly defines either of these data sources as a measure of mechanisms, i.e. not as a main outcome.

Intervention monitoring. Sign-in sheets will be used to track teachers' exposure to capacity-building activities. Avanti Fellows will moreover provide data from its software backend, to track teachers' use of videos and digital learning materials. Monitoring data will also be collected in the above-mentioned bi-weekly and monthly visits, through a structured school questionnaire.

Cost data. Avanti Fellows will provide data on program implementation costs (planned and actual).

4 Analytic strategy

4.1 Primary outcomes

The study's main focus is the extent to which student learning improves, because of the program. In summary, to investigate this question, I will compare students' performance across the three groups of schools, along with schools' intended or "intent-to-treat" status.[14]

More technically, I will estimate the intent-to-treat effect (ITT) of the two different treatments on endline learning outcomes, using the following specification:

$$Y_{irs2} = \alpha_s + \phi_r + Y_{irs1} + \sum_{k=1}^{2} \beta_{ks} T_{ki} + \boldsymbol{X}_{irs1} + \epsilon_{irs2} \tag{1}$$

[13]This work greatly benefited from conversations with experts on classroom observation measures (in particular, Professors Heather Hill of Harvard and Edward Seidman of NYU). I thank the World Bank's SABER team for sharing its newly developed "TEACH" classroom observation instrument with me. In summary, the study's instrument measures six dimensions, as follows: monitoring of student learning; feedback; maximization of learning time; density of the mathematics / science; clarity of content and lack of errors; richness of the mathematics / science.

[14]Further below, I will describe additional analyses that relate program effects to actual, observed program implementation.

, where Y_{irst} is the outcome of interest (standardized to $\mu = 0$ and $\sigma = 1$, on the baseline test), for student i, in randomization stratum r, and subject s, at period t ($t = 1$ denotes baseline; $t = 2$ denotes endline). T_k is the dummy for treatment k. $\boldsymbol{X}_{irs1}$ is a vector of student covariates measured at baseline[15]; ϕ_r are randomization strata fixed effects and ϵ_{irs2} captures the idiosyncratic error term (expected to be correlated at the school-level (cf. Abadie et al. 2017)).

I will thus assess the following research hypotheses, by testing their corresponding null, H.

1. The two variants of the program affect student learning in subject s. H_1: in Equation 1, $\beta_{ks} \neq 0$

2. The two variants of the program affect student learning in subject s *differently*. H_2: in Equation 1, $\beta_{1s} \neq \beta_{2s}$

This study's analyses will rely on Randomization Inference (RI) (cf. Young 2016). Athey and Imbens (2017) discuss how pairwise randomization may complicate the econometric analysis of randomized trials (e.g. for regression-based methods). Moreover, the authors (ibid.) and Banerjee et al. (2017) discuss additional complications due to re-randomization, if a given analytic strategy does not consider which feasible randomizations have been ruled out. Therefore, I will prefer an RI-based analysis of research hypotheses, permuting schools' treatment status within the given strata, and mimicking the above-mentioned (re-)randomization procedure 5,000 times (per specification).

4.2 Effects on subskills

To provide a more fine-grained understanding of program effects, I will repeat the above analyses by subskill. For mathematics, items are categorized as measures of either one of the following four abilities: Algebra; geometry; number system; and data/statistics. For science, items are categorized as measures of either biology, chemistry, or physics. In summary, this subskill analysis will establish the extent to which the program affected a student's probability of mastering each of these subskills.

In the case of subskills, because of the lower number of items per sub-skill, student performance will not be scored with a continuous measure of ability. Instead, I will distinguish between three categories of mastery, for each of the two subjects: students who have mastered grade-level appropriate material; students who have only mastered material from one and two

[15]This vector will include a student's gender, her age, and her grade-level, at baseline. In addition, I will include village- and school-level covariates, from the 2011 census and from administrative data; in doing so, I will follow Dhar et al. (2018) and select these additional controls using Lasso.

years below their grade-level; and other students. I will determine students' level of mastery empirically, through a Cognitive Diagnostic Model (CDM).[16] A short, technical appendix to this pre-analysis plan provides additional details on this CDM.

4.3 Heterogeneous effects

The program may affect some student subgroups more strongly than others; in addition, it can be expected that the program will increase its effectiveness after its initial inception phase. I will investigate three types of such heterogeneity in program effects. To avoid specification searching, I pre-specify these analyses for subgroups here, as follows.[17]

By initial ability level (at baseline). I will distinguish between three categories of mastery, for each of the two subjects. To determine students' level of mastery, I will employ a Cognitive Diagnostic Model, as described above.

By grade level. I expect heterogeneous effects by students' enrolled grade-level, at baseline (grade 9, grade 10).

By cohort. I expect effects to be larger in the second cohort of the program, once program implementation has stabilized.

Heterogeneous effects posit an interaction between the treatments' indicator variables with the respective baseline characteristic. For illustration, I provide the corresponding equation for heterogeneous effects by grade-level, as follows:

$$Y_{irs2} = \alpha_s + \phi_r + Y_{irs1} + \sum_{k=1}^{2} \beta_{ks} T_{ki} + \sum_{k=1}^{2} \beta_{2+ks} T_{ki} * G_{irs1} + \boldsymbol{X}_{irs1} + \epsilon_{irs2} \tag{2}$$

, where G represents a binary indicator of whether a student was enrolled in grade 10, at baseline. More formally, the study thus hypothesizes that the program's two variations affect student learning in subject s differently depending on a student's sub-category status, at baseline. H_3: in Equation 2, $\beta_{2+ks} = 0$. In the given illustration, this coefficient will indicate whether the program affects tenth-grade students more strongly (or not as much), in comparison to ninth-grade students.

[16] To determine levels of mastery, I will thus not use arbitrary cut-offs, such as whether students answered 50 percent (or any other percentage) of items correctly.

[17] Specification searching or "p-hacking" refers to occasions in which "a researcher, consciously or not, adjusts his model specifications or analysis sample in numerous ways until he finds a significant coefficient on the explanatory variable(s) of interest." (Tanner 2018, 1). My analyses will not consider any additional heterogeneous effects, beyond the ones specified just below.

4.4 Secondary outcomes: Analysis of mechanisms

What are the intermediary, secondary outcomes that drive the program's effects on student learning? To answer this question, I will turn to an analysis of mechanisms. I will approach this issue in two steps. First, I will assess whether teaching behaviors and instructional quality change because of the program. To this end, I will replace Y_{irs2} in Equation 1 with the above-mentioned measures (a) of teaching behaviors and (b) of instructional quality. Second, I will explore the association of each of the instructional measures with the continuous measures of student ability, in math and science. Finally, following Romero et al. (2017), I will report on the full mediating pathway, separating direct program effects on student learning from effects that are channeled through changes in instructional behaviors and instructional quality.[18]

4.5 Additional analyses

I will moreover conduct the following analyses.

1. Assessment of internal validity and robustness of results, in particular:
   (a) Analysis of sample balance on baseline observables, across study groups
   (b) Analysis of (un-)systematic attrition, across study groups
   (c) Robustness to handling of missing baseline information, by re-estimating Equations 1 and 2 without baseline covariates (Y_{irs1} and $\boldsymbol{X}_{irs1}$)
   (d) Robustness to adjustments that account for multiple hypothesis testing, following Young (2016)[19]
2. Instrumental Variable (IV) estimates of dose-response relationship (cf. Muralidharan et al. 2016, section 4.5)
3. Comparison of either program variant's cost-effectiveness (cf. Dhaliwal et al. 2014)

[18] Romero et al. (2017) employ a Lasso procedure to identify potential mediators (from a large number of variables). I will follow the authors' approach (*ibid.*) to identify mediating instructional behaviors; yet, the Lasso procedure will be forced to include teachers' use of ICT materials as a predictor. In turn, concerning instructional quality, I pre-specify six dimensions that will be investigated as potential mediators (monitoring of student learning; feedback; maximization of learning time; density of the mathematics / science; clarity of content and lack of errors; richness of the mathematics / science). This approach to mediation analysis rests on sequential ignorability – hence, I will interpret these results conservatively (see Romero et al. 2017).

[19] In the case of main program outcomes (i.e. continuous measures of student learning, in math and science), I will not use these adjustments.

5 Reporting

The study will produce two types of reports: (a) internal reports to Avanti and GoH on student performance, for each round of assessments, and (b) reports on the academic findings of the study.[20]

1. **Internal reports on student performance**: I will accompany each round of assessments with a report on student performance. Aggregated to the school-, block-, and district levels, these reports will present average scores (normalized), percent of items answered correctly, and percentage of students having mastered a grade-level understanding of either subject.

2. **Academic reports**: Academic reports will strictly focus on the program's effects, following this pre-analysis plan. Prior to January 1, 2022, academic working papers and presentations will be shared with HSSPP and academia for discussion of interim findings. After January 1, 2022, reports and materials will be published. GoH and Avanti will be clearly acknowledged in any of these publications and materials.

[20] In addition, Avanti will provide half-yearly reports to the HSSPP.

Appendix A Technical appendix

This brief technical appendix provides additional information on how students' ability will be categorized across three levels of mastery, using Cognitive Diagnosis Models (CDMs). CDMs are multi-dimensional latent-trait models, which were "developed specifically for diagnosing the presence or absence of multiple fine-grained skills or processes required for solving problems on a test" (de la Torre 2009, 164). This study will largely rely on the generalized deterministic inputs, noisy and gate (G-DINA) model for dichotomous items (de la Torre 2011). As common for CDMs, the G-DINA model requires a theoretically-founded specification of which attributes are expected to contribute to an examinee's probability of answering a given item j correctly. This so-called "Q-matrix" lists all items as rows, all attributes as columns, and denotes $q_{ja} = 1$ if attribute a is reflected in item j (and $q_{ja} = 0$, otherwise). The study's student assessments are explicitly designed to provide this item-to-skill mapping.

In CDMs, the mastery profile of each learner is described by a latent vector of dichotomous entries that each indicate whether an examinee has mastered any attribute; $\boldsymbol{\alpha}^*_{lj} = (\alpha_{l1}, \cdots, \alpha_{lk}, \cdots, \alpha_{lK^*_j})$, where K^*_j denotes the number of attributes captured by item j. Conditional on this latent vector $\boldsymbol{\alpha}^*_{lj}$, G-DINA models the probability of an examinee's correct answer for j, as a function of item parameters λ_j.

Following de la Torre (2011), we may express a respondent's probability of solving an item as

$$P(X_j = 1|\boldsymbol{\alpha^*_{lj}}) = \lambda_{j0} + \sum_{k=1}^{K^*_j} \lambda_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K^*_j} \sum_{k=1}^{K^*_j-1} \lambda_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \lambda_{j12\ldots K^*_j} \prod_{k=1}^{K^*_j} \alpha_{lk} \qquad (3)$$

, where λ_{j0} reflects the probability of a correct answer to item j for non-masters (the "guessing parameter"), λ_{jk} is the main effect related to having mastered attribute k, $\lambda_{jkk'}$ captures the interaction effect for attributes k and k', and $\lambda_{12\ldots K^*_j}$ is the interaction effect given mastery of attributes 1 to K^*_J.

After an initial item screening, I will begin my analyses with a (saturated) G-DINA model, using marginal maximum likelihood estimation. Recall that I intend to measure student ability on two scales – one that reflects mastery at a student's enrolled grade-level, and one that reflects mastery at a grade-level below. Therefore, I perform this and any of the remaining estimations in eight runs (one per ability level, grade-level, and subject combination), allowing item parameters to vary. I then investigate whether model fit could be improved by using a log-linear or logit link instead of the above-mentioned identity link. I subsequently validate and refine the study's Q-matrix, following Torre and Chiu (2016), and through a qualitative review. Thereafter, following Ma et al. (2016), I investigate whether a more parsimonious,

reduced model may be used without a significant loss in model data fit, for each of the test's items. Lastly, after this model-selection procedure, I estimate students' mastery profiles.

Appendix B Supplementary tables and figures

Table 1: Program components, partner responsibilities, and intervention groups

	Avanti Fellows Role	**Role for GoH, and its subsidiary, the Haryana School Shiksha Pariyojna Parishad (HSSPP)**	**80 Government Senior Secondary Schools (GSSS) randomly assigned to ICT Group (Group A)**	**80 GSSS randomly assigned to Workbook Group (Group B)**	**80 GSSS randomly assigned to Control Group (Group C)**
Capacity Building Workshops for Teachers	Training design and implementation	Provision of master trainers from the District Institute of Educational and Training (DIET)	Provided to all teachers	Provided to all teachers	No training
ICT infrastructure	NIL	HSSPP to purchase install, and maintain projectors, computers and sound systems	2 Classrooms set up per school	No ICT infrastructure	No ICT infrastructure
Workbooks	Avanti to design and provide pdfs for printing	HSSPP to print and distribute workbooks	Workbooks provided to all students	Workbooks provided to all students	No Workbooks provided
Assessments	Design and provision of test papers and OMR pdfs to HSSPP for printing. Support in invigilation and spot checks	HSSPP to print, distribute and conduct the test	Baseline and Endline Tests conducted	Baseline and Endline Tests conducted	Baseline and Endline Tests conducted
Classroom Observation	Monthly observation by Avanti Program Managers (1-2 Classroom observation per school per month)	Monthly Observation by Master Trainers (2 Classroom Observations per school per month)	Monthly observation (1-2 Classroom Observations per school per month)	Monthly observation (1-2 Classroom Observations per school per month)	Monthly observation (1 Classroom Observation per school per month)

Notes: ICT: Information and communication technology. GoH: Government of Haryana. DIET: District Institute of Education and Training. HSSPP: Haryana School Shiksha Pariyojna Parishad. GSSS: Government Senior Secondary School.

References

Abadie, A., S. Athey, G. Imbens, and J. Wooldridge (2017, November). When Should You Adjust Standard Errors for Clustering? Technical Report w24003, National Bureau of Economic Research, Cambridge, MA.

Arancibia, V., A. Popova, and D. K. Evans (2016). Training Teachers on the Job: What Works and How to Measure it. Working Paper 7834, The World Bank, Washington, D.C.

Athey, S. and G. Imbens (2017). The Econometrics of Randomized Experiments. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017, November). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A., S. Chassang, and E. Snowberg (2017). Decision Theoretic Approaches to Experiment Design and External Validity. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. *Statistical Theories of Mental Test Scores*.

Bruhn, M. and D. McKenzie (2009, October). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

Carril, A. (2016). Dealing with misfits in random treatment assignment.

de la Torre, J. (2009, May). A Cognitive Diagnosis Model for Cognitively Based Multiple-Choice Options. *Applied Psychological Measurement 33*(3), 163–183.

de la Torre, J. (2011, April). The Generalized DINA Model Framework. *Psychometrika 76*(2), 179–199.

Dhaliwal, I., E. Duflo, R. Glennerster, and C. Tulloch (2014). Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries - A General Framework with Applications for Education. In P. Glewwe (Ed.), *Education Policy in Developing Countries*, pp. 285–338. Chicago: The University of Chicago Press.

Dhar, D., T. Jain, and S. Jayachandran (2018, June). Reshaping AdolescentsâĂŹ Gender Attitudes: Evidence from a School-Based Experiment in India.

Egert, F., R. G. Fukkink, and A. G. Eckhardt (2018, January). Impact of In-Service Professional Development Programs for Early Childhood Teachers on Quality Ratings and Child Outcomes: A Meta-Analysis. *Review of Educational Research* (Online-first).

Ganimian, A. J., A. de Barros, and K. Muralidharan (2017, September). Do Students Benefit from Personalized Learning? Experimental Evidence from India.

Glennerster, R. (2017). The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Economic Field Experiments*, Volume 1, pp. 175–243. Elsevier.

Gove, A., M. K. Poole, and B. Piper (2017, March). Designing for Scale: Reflections on Rolling Out Reading Improvement in Kenya and Liberia. *New Directions for Child and Adolescent Development 2017*(155), 77–95.

Jacob, B. and J. Rothstein (2016, September). The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives 30*(3), 85–108.

Kolen, M. J. and R. L. Brennan (2004). *Test Equating, Scaling, and Linking* (3rd ed.). New York, NY: Springer.

Kraft, M. A., D. Blazar, and D. Hogan (2018, February). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, 0034654318759268.

Ma, W., C. Iaconangelo, and J. de la Torre (2016, May). Model Similarity, Model Selection, and Attribute Classification. *Applied Psychological Measurement 40*(3), 200–217.

Muralidharan, K. and P. Niehaus (2017, November). Experimentation at Scale. *Journal of Economic Perspectives 31*(4), 103–124.

Muralidharan, K., A. Singh, and A. J. Ganimian (2016, December). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. Working Paper 22923, National Bureau of Economic Research.

Romero, M., J. Sandefur, and W. A. Sandholtz (2017, September). Can Outsourcing Improve Liberia's Schools? Preliminary Results from Year One of a Three-Year Randomized Evaluation of Partnership Schools for Liberia. Working Paper 462, Center for Global Development, Washington, D.C.

Samejima, F. (1973). A Comment on Birnbaum's Three-Parameter Logistic Model in the Latent Trait Theory. *Psychometrika 38*(2), 221–233.

Stocking, M. L. and F. M. Lord (1983, April). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement 7*(2), 201–210.

Tanner, S. (2018). Evidence of False Positives in Research Clearinghouses and Influential Journals: An Application of P-Curve to Policy Research.

Torre, J. d. l. and C.-Y. Chiu (2016, June). A General Method of Empirical Q-matrix Validation. *Psychometrika 81*(2), 253–273.

Vivalt, E. (2017, September). *How Much Can We Generalize From Impact Evaluations?* Job market paper, Australian National University, Canberra, Australia.

Young, A. (2016, February). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results.