The Long Shadow of Early Education: Evidence from a Natural Experiment in the Philippines*

Thomas Lloyd Dean Yang

July 15, 2024

Abstract

How does early educational quality affect long-term academic outcomes? We exploit a natural experiment in the Philippines—a flawed implementation of mother-tongue education policy in public schools for kindergarten through Grade 3 starting in 2012—to examine this question. This policy led to an unexpected decline in educational quality, affecting only a subset of schools strongly predicted by pre-policy student language composition. We use language composition variables as instrumental variables for treatment. Leveraging panel data and confirming robustness to pre-trends, we find that the policy implementation: 1) had null effects on Grade 3 test scores, 2) led to declines in Grade 6 test scores across all subjects, and 3) reduced student enrollment and teacher retention in public primary schools. Employing a triple-difference strategy with Philippine Census data (across cohorts, localities, and 2010/2020 censuses), we show that by 2020, younger cohorts in highly-exposed localities completed 0.3 fewer years of schooling. Our findings demonstrate the substantial and enduring impact of early education quality on later academic achievement, contributing to the literature on human capital formation and education policy effectiveness.

Keywords: education quality, mother-tongue education, natural experiment, long-term educational outcomes, Philippines, human capital formation
JEL Codes: I21, I28, O15, C26, H75

*Thomas Lloyd: University of Michigan (tmlloyd@umich.edu); Dean Yang: University of Michigan, NBER and BREAD (deanyang@umich.edu). Yu Shi and Jiaxin Guo provided exemplary research assistance. We appreciate thoughtful feedback from Kevin Carney, Taryn Dinkelman, Susan Godlonton, Brian A. Jacob, and Yusuf Neggers, as well as seminar participants at the Causal Inference in Education Research Seminar (CIERS), the University of Michigan Development Workshop, and the University of Michigan 2024 MDev Conference. We thank DepEd officials at the Policy Research and Development Division (PRDD), the Education Management Information System Division (EMISD), the Planning Service (PS), the Bureau of Education Assessment (BEA), the Bureau of Learning Resources (BLR), the Bureau of Learning Delivery (BLD), Roger Masapol, former Director of Planning Services, Noel T. Baluyan, Assistant Secretary for Administration, and Gina O. Gonong, Undersecretary for Curriculum and Teaching (CT) for sharing data and coordinating this study. We also thank the staff at Innovations for Poverty Action (IPA) Philippines for field support and notably Karisha Anne Cruz for excellent project management.

1 Introduction

The quality of early education is widely believed to play a crucial role in shaping long-term academic and economic outcomes. Prior research has demonstrated positive correlations between early educational experiences and later life success (e.g., Heckman (2006); Chetty et al. (2011)). However, our understanding of the impact of early education quality remains incomplete. While some studies have shown short-term benefits of high-quality early education programs (Schweinhart et al., 2005), evidence on the longevity of these effects is mixed, with some research suggesting fade-out of initial gains (Bailey et al., 2020). Much of the existing literature focuses on specific interventions, leaving open questions about the impact of broader changes in educational quality. Causal evidence on how mainstream early education quality affects long-term outcomes is particularly scarce, especially in developing countries. These knowledge gaps hinder our ability to design effective educational policies and allocate resources optimally.

Empirically isolating the causal impact of early educational quality on later-life outcomes remains a significant challenge. Family background, socioeconomic status, peer effects, and subsequent educational experiences all shape outcomes, making it difficult to disentangle the specific contribution of early education quality. Selection issues and reverse causality further complicate analysis. Experimental studies, while providing strong internal validity, are typically limited in scale and duration. The gradual nature of many educational policy changes also makes it challenging to identify exogenous shocks facilitating causal inference. Consequently, our understanding of the causal relationship between early education quality and long-term outcomes remains incomplete, particularly for broad, system-wide changes.

To address these challenges, our study leverages a unique natural experiment. In 2012, the Philippine government implemented a mother-tongue education policy in public schools for kindergarten through Grade 3. While well-intentioned, the policy's flawed implementation led to an unexpected decline in educational quality. The switch in the medium of instruction – the "treatment" – occurred only in a subset of schools, which can reliably be predicted by pre-policy student language composition. This variation creates a quasi-experimental setting that allows us to overcome many of the identification challenges outlined above. We use pre-policy student language composition measures as instrumental variables for "treatment" (i.e., exposure to lower quality early education) at the school level, addressing potential selection issues. In addition, the unexpected nature of the quality decline mitigates concerns about reverse causality. Furthermore, our access to rich panel data enables us to confirm robustness to pre-trends, strengthening our identification strategy. By combining this natural experiment with a triple-difference approach using census data, we can trace the impacts of this early educational shock through to later academic performance and overall educational attainment. This methodology allows us to provide causal estimates of mainstream early education quality affects long-term outcomes.

Our empirical analysis yields several key findings. First, we find that exposure to the flawed mother-tongue education policy implementation had null effects on Grade 3 test scores, and substantial negative effects on Grade 6 test scores across all subjects: treated students score 0.67 standard deviations lower on Grade 6 tests three years after exposure to the policy. Second, we observe notable declines in student enrollment and teacher retention in treated public schools. Third, our analysis of census data reveals long-term effects on educational attainment. Our triple-difference estimates indicate that by 2020, younger cohorts in high-treatment areas completed 0.3 fewer years of schooling compared to their peers in low-treatment areas. This result is both statistically significant and economically meaningful. Collectively, these findings provide causal evidence of the enduring consequences of early education quality on academic achievement and educational progression over the longer term.

Our work is related to a body of prior research. The importance of early childhood education quality has been extensively documented in the literature. Seminal work by Heckman (2006) emphasizes the critical role of early investments in human capital formation. A broad set of studies has shown the existence of "sensitive periods" or "critical periods" – stages in life where health, economic, social, or other conditions have a persistent impact on later-life outcomes (Cunha et al. (2006), Almond and Currie (2011), and Currie and Almond (2011)). Studies such as the Perry Preschool Project (Schweinhart et al., 2005) and evaluations of Head Start (Ludwig and Miller, 2007) have demonstrated positive short-term effects of high-quality early education programs. However, the persistence of these effects remains debated, with some research suggesting fade-out of initial gains (Bailey et al., 2020), while others find enduring impacts (Chetty et al., 2011).

Our work also contributes to the literature on education in developing countries that highlights unique challenges and policy considerations. Glewwe and Muralidharan (2016) emphasize the need for context-specific research and policy solutions on the economics of education in developing countries. In recent decades, developing countries have experienced a large expansion of schooling, with the average years of formal education more than tripling from 1950 to 2010 (Barro and Lee, 2013).[1] However, such gains in years of education do not always translate into learning gains or human capital gains (Pritchett, 2013; World Bank, 2018; Muralidharan et al., 2019). Because other studies do document the potential for schooling have large returns (Duflo, 2001), prior research has studied the potential explanations for the inefficiency of schooling in developing countries. Common candidate explanations for inefficiency include low levels of spending associated with shortages in teaching materials and staff, over-ambitious or unadapted curricula (Banerjee et al., 2016; Muralidharan et al., 2019) with students who fall behind never given the opportunity to catch up, and teacher absenteeism associated with weak teacher incentives (Kremer et al., 2013; Mbiti et al., 2018).[2]

We also provide novel insights on the impacts of mother tongue education policies in multilingual contexts. A number of studies have explored the role of the language of instruction in the human capital production function (Angrist and Lavy, 1997; Angrist et al., 2008; Argaw, 2016; Taylor and von Fintel, 2016; Ramachandran, 2017; Laitin et al., 2019). Mother tongue education policies are often motivated by the following causal chain: learning in the mother tongue may facilitate the acquisition of cognitive skills (both reading and numeracy skills) in early grades which may in turn improve the learning of a second language and the translation and expansion of such acquired skills in the second, dominant language (Taylor and von Fintel, 2016). It is such human capital gains in the second language that are expected to have the largest economic returns. The second link of this causal chain is the most controversial, namely the translation of skills into a second (dominant) language. Opponents worry that mother tongue instruction may actually *reduce* proficiency in the dominant language.[3] However, another important link, upstream in the causal chain, that

[1] From 2.0 years in 1950 to 7.2 years in 2010 (for those aged 15 and over). The worldwide average years of schooling for those aged 15 and above is 7.9 years.

[2] The Philippines ranks the lowest in expenditure per student among participants in the PISA 2018 survey, and 90% lower than the OECD average (OECD, 2018).

[3] There is currently mixed evidence on this issue. For example, using quasi-random variation in Ethiopia, Argaw (2016) finds that mother tongue-based education leads to a 11 p.p. gains in reading skills and modest gains in labor market outcomes. In contrast, using a randomized experiment in Kenya, Piper et al. (2018) find no effect of mother tongue instruction on literacy skills in English

is often overlooked, relates to the feasibility of teaching in the mother tongue and to the potential shock to education quality associated with a shift to instruction in local languages without adequate preparation. Bühmann and Trudell (2008) argue for the benefits of mother-tongue education in improving learning outcomes, while Heugh (2012) highlights challenges in implementing such policies in developing countries. The complex linguistic landscape of the Philippines, as described by Tupas and Martin (2017), provides a relevant context for examining these issues.

Finally, this paper also highlights challenges of policy implementation at scale (Angrist et al., 2023; Angrist and Meager, 2023; List, 2022). Pritchett et al. (2013) discuss the complexities of policy implementation in developing countries, while Bold et al. (2018) provide evidence on how well-intentioned educational interventions can fail to deliver expected results at scale. Our work documents that a mother tongue education policy in a multilingual context, implemented nationwide—and associated with important implementation challenges—led to a sharp reduction in test scores and longer-run educational attainment.

Our study contributes to these strands of literature by leveraging a unique natural experiment in the Philippines to provide causal evidence on the long-term impacts of early education quality. By examining the effects of a broad, system-wide change in educational quality, rather than a targeted intervention, we address a significant gap in the literature. Furthermore, our focus on a developing country context and our ability to trace impacts over several years adds valuable insights to our understanding of human capital formation and the persistence of early educational effects.

2 Context: Mother tongue education in the Philippines

The Philippines is a highly linguistically diverse country with a total of 184 distinct spoken languages reported in the Ethnologue (Ebernhard et al., 2023), and 245 distinct languages and dialects reported in the 2020 Census. Tagalog is the most widely spoken language, with 34.0% of primary school students declaring it as their mother tongue, closely followed by Cebuano/Bisaya/Binisaya at 25.3%.[4] In 1973, in an effort to reconcile the Philippines' colonial history with its postcolonial nation-building objectives, as well as ethnolinguistic ideologies, the country adopted a bilingual ed-

and slightly negative impacts on numeracy skills.

[4] Other notable language groupings include Hiligaynon/Ilonggo at 7.4%, Ilocano at 6.7%, and Bikol at 5.7%.

ucation system, with both English and Filipino (a standardized form of Tagalog) as languages of instruction (Tupas and Martin, 2017; Monje et al., 2019).

This bilingual system created the scope for a mismatch between a child's mother tongue and their school's language of instruction. Many students were induced to learn a second language the minute they set foot in school. Such language mismatch was associated with inequalities in access to learning in early childhood, stigma and marginalization. For this reason, language of instruction was identified by the Department of Education of the Philippines (DepEd) as a potential determinant for the country's relative poor performance in international large-scale assessments studies such as the Trends in International Mathematics and Science Study (TIMSS) in 1999 and 2003.(DepEd, 2009) Poor performance of the bilingual education system,[5] together with endorsements for the use of local languages from international organizations (Bühmann and Trudell, 2008; Ball, 2010) and a desire to reconnect with local cultural identities (Tupas and Martin, 2017), formed the rationale for shifting to mother tongue instruction.

In school year 2012-2013, the Department of Education of the Philippines implemented at a large scale (nationwide)—and without adequate preparation—a mother tongue-based education policy in early primary school (DepEd, 2012). The Mother Tongue-Based Multilingual Education (MTB-MLE) policy induced a switch in the medium of instruction (MOI) for schooling from Kindergarten to Grade 3. Instruction in English and Filipino, the two national languages, was replaced by instruction in the mother tongue. Filipino is the institutionalized version of Tagalog[6], the regional language spoken widely in the National Capital Region, Central Luzon, Calabarzon, and Mimaropa, among others. As a result, the policy affected the following two groups differently: (i) treated schools, i.e., schools that changed their medium of instruction to a language other than Tagalog post-policy, and (ii) control schools, i.e., schools whose medium of instruction was Tagalog pre- and post-policy.

Basic Education in the Philippines. As of school year 2020-2021, there were a total of 22.6 million students enrolled in public schools in K-12, of which 11.6 million were enrolled in elementary school (Grades 1 to 6). The education budget per student averages approximately \$514 per student per year as of 2022 (compared to ≈ \$15,000

[5] One argument from the 2009 DepEd memo reads as follows: "top performing countries in the Trends in International Mathematics and Science Study (TIMSS) are those that teach and test students in science and math in their own languages".

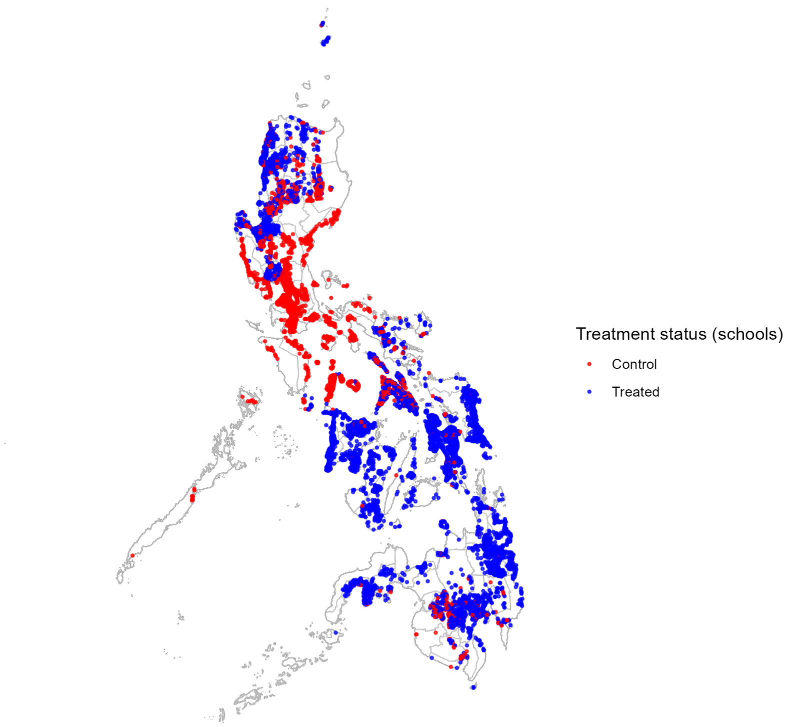
[6] Throughout the rest of the paper, we will use the term "Tagalog" only to avoid confusion.

in the U.S. (Hanson, 2024)). Grade completion is high, with 90.3% of those aged 9-10 completing Grade 3, 87.1% of those aged 12-13 completing Grade 6, and 68.0% of those aged 16-17 completing Grade 10 in the 2020 Census. At the same time, Filipino students exhibit poor performances in international large-scale assessments such as the 2018 PISA and the 2019 TIMSS (OECD, 2018; Mullis et al., 2020), and 9 out of 10 students at late primary age struggle to read and comprehend simple texts (World Bank, UNESCO, 2021).

The MTB-MLE policy. A total of 19 languages are officially recognized as MOI that schools can choose to teach in from K to G3 following the MTB-MLE guidelines. However, in practice, schools were recommended and encouraged to teach in the language that students know best i.e. to use learners' "First language (L1)" or their "native language" (DepEd, 2009). As a result, many schools teach in a local MOI that is not among this list of 19 languages to better cater to the local context and to their students' need. Moreover, and as can be seen from Figure 1, many schools choose to teach in Tagalog (and select into the control group) in provinces where Tagalog is not the most widely spoken language. Other than using the mother tongue as the language of instruction, schools were tasked with developing their own learning resources such as writing classroom materials on language, literature and culture, documenting the orthography and grammar of the language, and developing a dictionary of the language (Monje et al., 2019).

The sudden nationwide implementation of the MTB-MLE policy has been met with conceptual and implementation challenges. On the conceptual front, this "one-size-fits-all" policy does not account for linguistic diversity within the classroom, or for the breadth of local dialects of the same language (Monje et al., 2019). It also ignored the lack of standardization and intellectualization of the orthography and grammar of local languages needed for instruction (especially in mathematics and science) (Metila et al., 2016). On the implementation front, challenges include teachers' lack of proficiency to teach in a school's MOI, students who do not speak their school's MOI, insufficient teacher training, a dearth or absence of textbooks and learning materials in the local mother tongue, and the unpopular status of the mother tongue as MOI among students, parents, and teachers (Monje et al., 2019; Tupas and Martin, 2017; Metila et al., 2016).

Figure 1: School-level Treatment Variation Across Space



Note: This figure shows the geographical location of public schools in our sample (with medium of instruction information) across the Philippines, and color-coded based on treatment status. Treated schools are shaded in blue while control schools are shown in red. Light grey lines demarcate province borders.

3 Data and Summary Statistics

We combine detailed administrative datasets obtained from the Philippine Department of Education (DepEd) at both the school and individual level together with survey and census data from the Philippine Statistics Authority (PSA). We summarize data sources below. Additional details are provided in Appendix A.1. The summary statistics for key variables are shown in Table 1.

3.1 Grade 3 & Grade 6 National Achievement Test (NAT) Scores

Our first main outcomes are nationally standardized test scores, administered by DepEd, for repeated cross-sections of students in Grade 3 and Grade 6 from school year (SY) 2008-2009 to SY 2017-2018. We use a 10% random sample of the universe of test score results spanning our 10-year study period.[7] Tests subjects include English, Filipino, and Mathematics in Grade 3 and Grade 6, as well as Science, and History & Geography (referred to as "Hekasi") in Grade 6. An *Overall* test score is computed as a simple average across subjects. We restrict our main sample to focus exclusively on public schools with medium of instruction information.This includes 24,529 public schools in 1,482 municipalities in 84 provinces with approximately 2 million test scores.

3.2 Linguistic data (Mother Tongue & Medium of Instruction)

Linguistic data. Data on the mother tongue of the universe of elementary public school students was first collected by DepEd in SY 2012-2013 and is used to construct measures of the linguistic composition of each school's student body *before* the policy. Using data from *never treated* students who were in Grade 4, Grade 5, or Grade 6 in 2012-2013 (approximately 6 million of the 12 million elementary school students in 2012-2013), we compute, for each school, the percentage of students speaking each of the 19 languages offered as media of instruction.[8] This approach ensures that we are not constructing school-level linguistic composition data with cohorts (those in Grades 1, 2, or 3) whose composition may have been affected by the MTB-MLE policy in its first year of implementation. We exploit this *pre*-policy linguistic composition data to instrument for the school-level choice of the medium of instruction *post*-policy, which determines treatment status at the school level.

Medium of instruction. Information on the medium of instruction adopted by each school post policy originates from two complementary data sources for which the sample is limited to a subset of the universe of public schools in the Philippines. The first is a DepEd-conducted survey of schools in 2022 profiling the medium of instruction adopted post MTB-MLE policy and spanning 20,120 schools. The second is a 2018-19 survey of 15,916 schools conducted by Monje et al. (2019) in their process evaluation study of the MTB-MLE policy implementation. We assign a school to the control group if it reported its medium of instruction to be Tagalog in either one of

[7] The sample was obtained via stratified random sampling on region, and school division conducted by the Bureau of Education Assessment at DepEd.

[8] Note that the correlation coefficients between the school-level linguistic variables constructed with all elementary grades (Gr 1 to Gr 6) and those constructed with only never treated grades (Gr 4 to Gr 6) are very close to 1. We use Gr 4 to Gr 6 students only to alleviate concerns about potential linguistic compositional changes for students in Gr 1 to Gr 3 in 2012-2013.

these two surveys.

3.3 Elementary Enrollment, Teachers, and Pupils-to-Teacher Ratio

We also examine school-level administrative data on student enrollment in each grade from Grade 1 to Grade 6, and total elementary teacher count provided by DepEd from SY 2008-2009 to SY 2017-2018. This allows the analysis of potential compositional changes resulting from the MTB-MLE policy both between schools and between grades, as well as the potential movements of teachers from treated to control schools (or vice versa), or the potential departure of teachers.

3.4 Grade Completion & Years of Education

We use data from the 2010 and 2020 Decennial Censuses of Population and Housing (CPH) collected by the Philippines Statistics Authority (PSA) covering the entire Filipino population to study learning outcomes such as grade completion and the number of completed years of education eight years after implementation of the policy. More specifically, we match data from approximately 35 million respondents aged 7 to 25 per census round with information on municipality of birth, municipality of residence, highest grade completed, and the age of the respondent, with DepEd school-level data aggregated up at the municipality level (used to define treatment intensity).

4 Empirical Analyses

We aim to shed light on the impacts of the MTB-MLE policy "treatment" (switching medium of instruction to a language other than Tagalog) on a variety of education-related outcomes. First, we examine outcomes using DepEd administrative data for which treatment status is determined at the school level, such as Grade 3 and Grade 6 test scores, Grade 1 to Grade 6 enrollment, elementary teacher counts and the pupils-to-teacher ratio (a proxy for class size). Second, in analyses of census data for which treatment intensity is defined at the municipality level, we examine impacts on respondents' highest grade completed.

Table 1: Summary statistics

	Mean	SD	N
Student-level variables SY 2008-09 to SY 2017-18			
Grade 3 Overall Test Scores	0.059	0.779	856,735
Grade 6 Overall Test Scores	0.061	0.751	1,102,850
School-level variables SY 2008-09 to SY 2017-18			
Grade 1 Enrollment Count	66.75	97.6	241,583
Grade 2 Enrollment Count	62.14	90.4	239,470
Grade 3 Enrollment Count	61.04	89.4	241,583
Grade 4 Enrollment Count	59.73	88.2	241,583
Grade 5 Enrollment Count	58.58	87.1	238,665
Grade 6 Enrollment Count	55.69	84.2	238,665
Number of Elementary Teachers	11.29	14.36	239,217
Elementary Pupils-to-Teachers Ratio (PTR)	33.71	10.71	236,585
School-level variables SY 2012-13			
Treatment status (Treat$_s$)	0.665	0.472	24,529
Pct. Tagalog (G1-G6) in 2012-2013	0.240	0.388	24,529
Pct. Tagalog (G4-G6) in 2012-2013	0.238	0.387	24,529
Census respondent-level variables (aged 7 to 25)			
Highest Grade Completed	7.274	3.264	73,267,484
Treatment intensity at the municipality level (Treat$_m$)	0.541	0.420	73,267,484

Note: This table shows summary statistics (sample mean, standard deviation, and the number of observations) for individual-level and school-level outcomes from DepEd administrative data used in our main analysis, as well as respondent-level outcomes from the 2010 and 2020 census rounds (for respondents aged 7 to 25). The sample is restricted to public schools with medium of instruction information. Treatment status (Treat$_s$) is a binary variable defined at the school level. A school is said to be treated if its medium of instruction post policy is **not** Tagalog (see Figure 1). Treatment intensity (Treat$_m$) is a continuous variable defined at the municipality level corresponding to the predicted percentage of treated students (see Figure 4).

4.1 Test Scores, Enrollment, and Pupils-to-Teachers ratio (DepEd data)

4.1.1 Empirical Approach

To estimate the causal effect of the MTB-MLE policy, our empirical strategy relies on a difference-in-differences approach. We start with the canonical two-way fixed effects (TWFE) *dynamic* specification in which we estimate the following regression equation allowing for differential treatment effects across time relative to 2011-2012 (baseline school year, $t = -1$):

$$Y_{ispt} = \alpha_s + \gamma_t + \eta_{pt} + \sum_{\substack{h=-4 \\ h\neq -1}}^{h=5} \tau_h \, \mathbf{1}\{t = h\} \times \text{Treat}_s + \varepsilon_{ispt}, \tag{1}$$

where Y_{ispt} is the outcome of individual i in school s, province p, and school year t. α_s and γ_t are school and school year fixed effects. η_{pt} are province-by-year fixed effects; their inclusion ensures that we rely exclusively on within-school variation over time between treated and control schools within the same province, corresponding to deviations from province-specific time effects. Treat_s is a binary variable equal to 1 if school s switched its medium of instruction to a language other than Tagalog post policy. τ_h are the parameters of interest corresponding to the average treatment on the treated units (ATT) in period h, with the absence of pre-trend hypothesis that $\tau_h = 0$ for $h < -1$, and period specific causal effects of the program $\tau_h > 0$ for $h \geq 0$. The leads $h \in \{-4, -3, -2, -1\}$ correspond to school years 2008-2009 to 2011-2012, while the lags $h \in \{0, 1, 2, 3, 4, 5\}$ correspond to school years 2012-2013 to 2017-2018. For lagged outcomes such as Grade 6 test scores (3-year lag allowing students to reach Grade 6), these indexes will be shifted by -3, with the baseline (omitted) school year shifting to 2014-2015 reducing the number of post periods to 3. Standard errors are clustered at the school level, the unit of treatment assignment. We also estimate a version of equation (1) aggregated up at the school level (suppressing the index i) for outcomes such as student enrollment and teacher counts.

Finally, we define $\tau_{\text{post}} = (1/\bar{T}) \sum_{h=0}^{\bar{T}} \tau_h$ as the average causal effect across all post-treatment periods.

Identification. τ_h are identified under the *parallel trend* and *no anticipation* assumptions. Moreover, our context satisfies the following three conditions highlighted by De Chaisemartin and D'Haultfoeuille (2023) which ensure unbiasedness for the ATTs: (i) the treatment is an absorbing state, (ii) the treatment is binary, and (iii) there is no variation in treatment timing. This avoids the problem of negative weights which could arise from comparing newly treated units relative to already treated units in designs with variation in treatment timing (Callaway and Sant'Anna, 2021; Borusyak et al., 2024).

Instrumental Variables (IV) approach. We augment the standard dynamic TWFE specification described by equation (1) with an instrumented difference-in-differences (IV-DID) approach. Although the school-level fixed effects (FEs) may address a part of endogeneity or selection concerns by controlling for all *time-invariant* school-level unobservables, it does not address potential selection on *time-variant* characteristics. In particular, we may worry that schools better able to teach in the local mother tongue select into treatment, and that schools less able to teach in the local mother tongue select out of treatment. To address these concerns, we instrument the binary school-level treatment indicator Treat_s (equal to 1 if a school switched to a medium of instruction other than Tagalog) with the percentage of SY 2012-2013 Grade 4 to Grade 6 learners (never treated cohorts) at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction, as well as a square and a cubic term in the percentage of Tagalog-speaking learners.

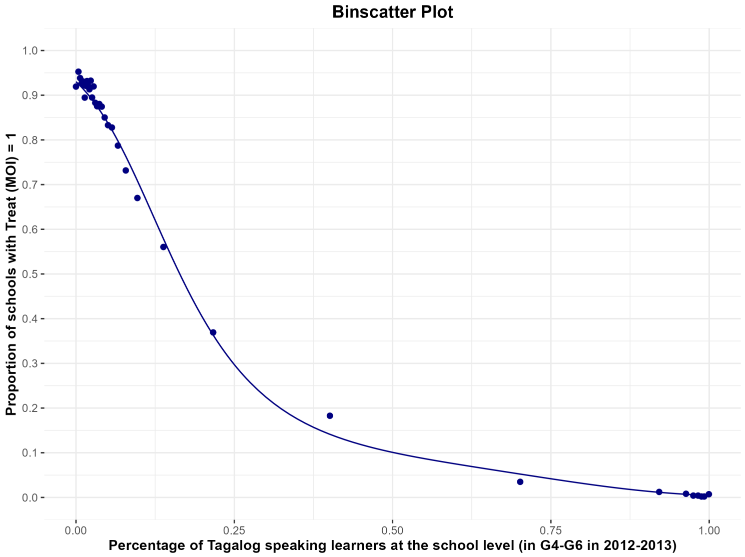
As discussed by Ye et al. (2023), the IV-DID approach is particularly attractive because it allows us to relax a key assumption for the standard IV. It is robust to violations of the exclusion restriction by allowing instruments to have a direct effect on the outcome. Instead, it only requires a weaker version of the exclusion restriction to hold: the instruments should have no direct impact on the *trend* in potential outcomes in the absence of treatment. Intuitively, in our context, as long as trends in outcome are parallel for schools with different values of our instrumental variables if all schools were counterfactually not exposed to the MTB-MLE policy, then any observed nonparallel trends in outcomes post-policy between schools provides evidence for a causal impact of the policy.

Finally, we also report an "honest" confidence interval in our main tables for the estimated average causal effect across post-treatment periods τ_{post} using the robust inference methods developed by Rambachan and Roth (2023) for difference-in-differences designs where the parallel trends assumption may be violated. More specifically, we use their "smoothness restrictions" approach on non-parallel trends in pre-treatment periods assuming no change in slope for the post-treatment periods (which corresponds to the case where $\bar{M} = 0$ using the notation from their paper) which intuitively is akin to controlling for a linear treatment group-specific time trend using only pre-treatment time periods. Intuitively, this method assumes that potential (linear) non-parallel trends would have persisted in the absence of the policy change and thereby adjust the coefficient estimates to capture significant breaks from these potential pre-trends. For example, a null estimated impact in the presence of a positive pre-trend may actually correspond to a non-negligible break in the pre-treatment trend.

4.1.2 First Stage

In this subsection, we present the results from the first stage of our IV estimation strategy. While in practice, for the IV coefficient estimates presented in the following

Figure 2: School-level treatment status and percentage of Tagalog-speaking learners



Note: This figure shows the binscatter plot together with a cubic fit illustrating the relationship between treatment status at the school-level (a school is treated if its medium of instruction post policy is not Tagalog) and the percentage of Tagalog-speaking learners in grades 4 to 6 during the 2012-2013 school year. The optimal number of bins and the cubic fit were generated using the data-driven approach described in Cattaneo et al. (2024) with a starting choice of $n = 50$ bins.

subsections, we instrument all the interactions between the treatment variable and the individual year dummy variables ($\mathbf{1}\{t = h\} \times \text{Treat}_s$) presented in equation (1) with the interactions between our full set of IVs and the year dummy variables, Table 2 shows the first stage results from the static analog for simplicity of exposure.[9] This table presents the coefficient estimates from the regression of Treat_s on all school-level linguistic composition variables used as instruments, and described above. Figure 2 shows non-parametrically, in a binscatter plot, the relationship between treatment status at the school level and the most predictive instrument, namely the percentage of Grade 4 to Grade 6 students (never treated cohorts) whose mother tongue is Tagalog in 2012-2013.

Strikingly, there is a very strong decreasing and convex relationship between the percentage of Tagalog-speaking learners pre policy and treatment status at the school level. The coefficient estimates in Table 2 confirms this graphical evidence, as indicated by the sign, magnitude and statistical significance of the linear, square, and cubic terms for the percentage of Tagalog-speaking learners.

The main takeaway from Table 2 is the very strong first stage with a F-statistic

[9] Note that linguistic variables are not collinear because students' mother tongue may be a language other than the 19 languages offered as media of instruction.

Table 2: School-level treatment status and linguistic composition pre-policy

	(1)		(2)		(3)
Pct. Tagalog	-2.205***	Pct. Hiligaynon	0.352***	Pct. Sambal	-0.090
	(0.098)		(0.015)		(0.129)
Pct. Tagalog sq.	2.367***	Pct. Waray	0.390***	Pct. Akeanon	0.405***
	(0.242)		(0.014)		(0.018)
Pct. Tagalog cb.	-0.796***	Pct. Tausug	0.120*	Pct. Kinaray-a	0.424***
	(0.154)		(0.062)		(0.015)
Pct. Cebuano	0.366***	Pct. Maguindanaoan	-0.283***	Pct. Yakan	0.361***
	(0.014)		(0.057)		(0.043)
Pct. Kapampangan	0.293***	Pct. Maranao	0.269***	Pct. Surigaonon	0.413***
	(0.021)		(0.025)		(0.017)
Pct. Pangasinan	0.425***	Pct. Chabacano	0.401***	Obs. (Schools)	24,529
	(0.021)		(0.019)	R^2	0.662
Pct. Ilocano	0.180***	Pct. Ibanag	-0.067	F	10,807.9
	(0.016)		(0.055)	Prob. > F	0.000
Pct. Bikol	0.282***	Pct. Ivatan	0.415***		
	(0.016)		(0.028)		

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the results from the estimation of a linear probability model (LPM) corresponding to our **first stage** equation in which we regress the binary (school-level) treatment indicator Treat_s (equal to 1 if a school switched to a medium of instruction other than Tagalog) on the percentage of SY 2012-2013 Grade 4 to Grade 6 learners (never treated cohorts) at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction, as well as a square and a cubic term in the percentage of Tagalog-speaking learners. These variables account for each school's student body linguistic composition pre policy. The binscatter plot in Figure 2 illustrates the (strongly predictive) decreasing relationship between treatment status and the percentage of Tagalog-speaking learners pre-policy.

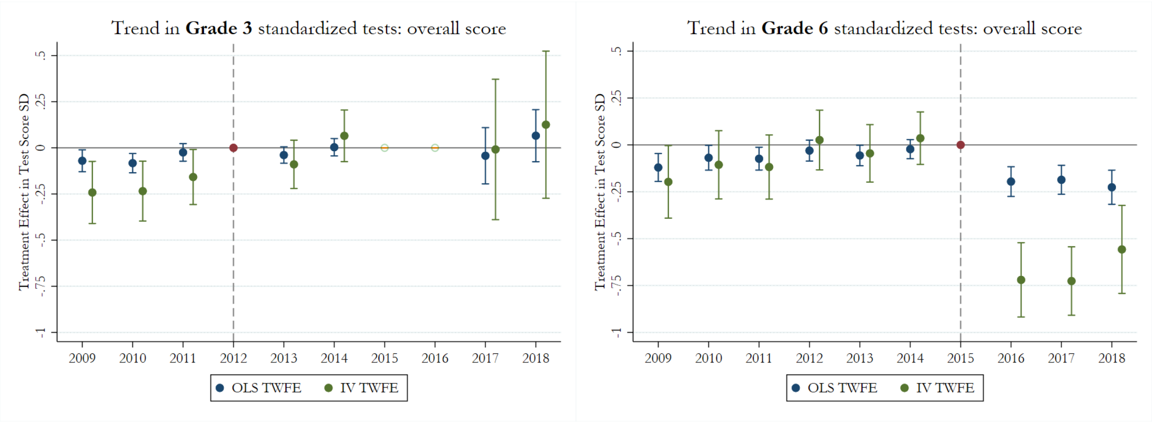
of 10,807 which empirically validates the assumption of trend relevance for our set of instruments. In this simple linear probability model, the linguistic composition variables explain 66.2% of the variation in treatment status. This suggests that when choosing whether or not to switch their medium of instruction, schools aimed to closely align their medium of instruction with the mother tongue of their students. Note that most coefficient estimates for the other languages offered as media of instruction are positive and statistically significant suggesting that a higher percentage of students speaking each of these languages increases the probability of treatment.[10]

[10] This is not the case for Ibanag nor Sambal which are estimated to have a null relationship with Treat_s, nor Manguindanaoan with a negative and statistically significant relationship.

4.1.3 Impacts on National Achievement Test Scores

Figure 3 presents the τ_h coefficient estimates on overall test scores from the estimation of equation 1, corresponding to the dynamic effects of the policy over time as well as the pre-trend tests. We show both OLS coefficient estimates (in blue) and IV coefficient estimates (in green). Panel (a) of Figure 3 shows the impacts on Grade 3 test scores from 2013 to 2018 and suggest a null impact of the policy in post treatment periods. Note however, the presence of a slightly positive pre-trend with students in treated schools on an slightly upward trajectory relative to those in control schools from 2009 to 2011. To account for this, we report an honest CI for τ_{post} in Table 3 using the robust inference tools developed by Rambachan and Roth (2023) for which we assume persistence of the pre treatment linear trend into post periods in the absence of the policy change. This 95% CI is considerably shifted leftward for Grade 3 overall test scores, and mostly negative with a midpoint of -0.14 SD but continues to include zero and positive values. This result for overall test scores is consistent across subjects in Grade 3 (see Table A1) with null coefficient estimates for English, Filipino, and Mathematics.

Figure 3: Dynamic Impacts on Grade 3 and Grade 6 *Overall* test scores



(a) Grade 3

(b) Grade 6

Note: Coefficient estimates (with 95% confidence intervals) from the estimation of equation (1) using the specification with school, school year, and province × year fixed effects. The dependent variable is Grade 3 *Overall* test scores in Panel (a), and Grade 6 *Overall* test scores in Panel (b). The pre-period is SY2008-2009 to SY2011-2012 for Grade 3 test scores, while it is SY2008-2009 to SY2014-2015 for Grade 6 test (accounting for a 3-year lag relative to SY2011-2012). IV estimates correspond to the instrumented DID specification where treatment status at the school level is instrumented with school-level linguistic composition variables pre-policy, i.e., the percentage of learners at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction as well as a square and cubic term in the percentage of Tagalog-speaking learners (first stage estimates are in Table 2). Standard errors are clustered at the school level.

We then turn to examining Grade 6 test scores, which measure longer-term learning once students transitioned back to the dominant language for instruction. As can be seen from Panel (b) of Figure 3, we find that the policy led to a sudden and substantial decline in Grade 6 overall test scores with a coefficient estimate of -0.67 SD for τ_{post} for our preferred IV specification (see column (2) in Table 3). As Table A2 shows, the decline in Grade 6 test scores holds across subjects. For example, treated students score 0.53 SD lower in mathematics, and 0.5 SD lower in science relative to students in control schools. The magnitudes of these coefficient estimates are quite large. In comparision, Evans and Yuan (2019) report that students learn between 0.15 and 0.21 standard deviation of literacy ability in a business-as-usual school year in a sample of low- and middle-income countries. If we extrapolate this to our setting, students are set back a little over three "equivalent years of schooling". Similarly, in a review of the literature, Evans and Yuan (2022) find average learning effect sizes of 0.15 SD across quasi-experimental studies (0.18 SD for reading, and 0.11 for mathematics) with the bottom percentile of -0.76 and a 90th percentile of 0.72.

Table 3: IV Estimates: Average Causal Effects on Grade 3 and Grade 6 *Overall* test scores

	Grade 3 Overall Score	Grade 6 Overall Score
	(1)	(2)
$\bar{\tau}_{\text{post}}$	0.023	-0.668***
	(0.086)	(0.093)
Honest CI (*smoothness restrictions*, $\bar{M} = 0$)	[-0.378, 0.101]	[-0.957, -0.532]
Control Mean, Pre Period	0.00	0.00
Year FE	Y	Y
School FE	Y	Y
Province × Year FE	Y	Y
Obs. (Students)	856,610	1,102,528
Clusters (Schools)	23,712	23,395

Note: Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the coefficient estimates on Grade 3 overall test score in column (1) and Grade 6 overall test score in column (2) for the average causal effect across post-treatment periods from the estimation of equation (1) using the instrumented DID specification where treatment status as the school level is instrumented with school-level linguistic composition variables pre-policy, i.e., the percentage of learners at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction as well as a square and cubic term in the percentage of Tagalog-speaking learners (see Table 2). The pre-period is SY2008-2009 to SY2011-2012 for Grade 3 test scores, while it is SY2008-2009 to SY2014-2015 for Grade 6 test (accounting for a 3-year lag relative to SY2011-2012). See Figure 3 for per period coefficient estimates.

Table A2 also reports the OLS coefficient estimates for τ_{post} across subjects. Likewise, these estimates are negative and statistically significant but smaller in magnitude than the IV estimates with a 0.2 SD decline in overall Grade 6 test scores. This difference between IV coefficient estimates and OLS coefficient estimates can also be seen in Panel (b) of Figure 3. This suggests the presence of positive selection into treatment, which leads the OLS coefficient to be positively biased ($\bar{\tau}^{\text{IV}}_{\text{post}} < \bar{\tau}^{\text{OLS}}_{\text{post}}$). Schools less able to teach in the local mother tongue may have selected out of treatment, biasing the OLS results in the positive direction.

4.1.4 Impacts on Enrollment

Table 4: IV Estimates: Average Causal Effects on Grade 1 to Grade 6 Enrollment Counts

	Gr 1	Gr 2	Gr 3	Gr 4	Gr 5	Gr 6
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{\tau}_{\text{post}}$	23.604***	2.942**	-3.391***	-7.079***	-7.174***	-6.890***
	(2.728)	(1.298)	(1.248)	(1.543)	(1.281)	(1.148)
Honest CI (*smoothness restrictions*, $\bar{M} = 0$)	[18.513, 35.496]	[2.748, 12.406]	[-5.177, 3.835]	[-6.507, 0.953]	[-8.116, -2.081]	[-7.390, -2.545]
Control Mean, Pre Period	101.3	85.3	81.6	78.5	77.0	73.9
Year FE	Y	Y	Y	Y	Y	Y
School FE	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y
Observations	241,572	239,459	241,572	241,572	238,654	238,654
Clusters (Schools)	24,527	24,527	24,527	24,527	24,234	24,234

Note: Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the coefficient estimates on grade-level enrollment from Grade 1 to Grade 6 for the average causal effect across post-treatment periods from the estimation of equation (1) using the instrumented DID specification where treatment status as the school level is instrumented with school-level linguistic composition variables pre-policy, i.e., the percentage of learners at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction as well as a square and cubic term in the percentage of Tagalog-speaking learners (see Table 2). The pre-period is SY2008-2009 to SY2011-2012 for Gr 1 to 3, it is shifted by 1 year for Gr 4, two years for Gr 5, and three years for Gr 6 (allowing time for treated students to reach these grades).

Next, we turn to grade by grade impacts on enrollment from Grades 1 to 6. Table 4 shows coefficient estimates for the average causal effect τ_{post} from our preferred IV specification while Table A3 reports OLS estimates. Strikingly, we find an interesting pattern across grades: enrollment in Grade 1 and Grade 2 increases considerably. This likely represents increased demand for instruction in the mother tongue in early grades. This may also reflect students increasingly repeating grades in early primary school. Moreover, enrollment significantly declines starting in Grade 3, and increasingly so in Grade 4, stabilizing in Grade 5 and Grade 6 with an estimated decrease in enrollment of 7 students per school and per grade corresponding to approximately 9%

of the control mean pre period in Grades 4, 5 and 6. This suggests that as students get exposed to instruction in the mother tongue and the implementation challenges associated with the policy, and progress through grades, they either repeat grades, transfer out to control schools, or drop out of school altogether.

4.1.5 Impacts on Teachers and the Pupils-to-Teacher Ratio (PTR)

The same way as students are fleeing treated schools, teachers may also be induced to leave the public sector or move to control schools as a result of the policy change. We test this hypothesis in this subsection.

Table 5 shows coefficient estimates for the average post-treatment causal effect τ_{post} for the Grades 1 to 6 teacher count and the Pupils-to-Teacher Ratio from the estimation of equation (1). Table A4 reports OLS estimates. Figure 7 presents the corresponding per period dynamic effects and shows a clear and statistically significant reduction in the number of primary school teachers post-policy, with up to 1 teacher leaving treated schools, on average. This corresponds to a 7.5% decline relative to the control mean from SY 2008-2009 to SY 2011-2012 of 13.26 teachers per school.

Table 5: IV Estimates: Average Causal Effects on Teacher counts and the PTR

	Teachers	PTR
	(1)	(2)
$\bar{\tau}_{\text{post}}$	-0.994***	0.379
	(0.192)	(0.644)
Honest CI (*smoothness restrictions*, $\bar{M} = 0$)	[-1.089, 0.100]	[-3.612, 1.966]
Control Mean, Pre Period	13.26	35.58
Year FE	Y	Y
School FE	Y	Y
Province × Year FE	Y	Y
Observations	239,140	236,494
Clusters (Schools)	24,461	24,431

Note: Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the coefficient estimates on elementary teacher count in column (1) and the Pupils-to-Teacher Ratio (PTR), for which values below the 1st and above the 99th percentiles were dropped, in column (2) for the average causal effect across post-treatment periods from the estimation of equation (1) using the instrumented DID specification where treatment status as the school level is instrumented with school-level linguistic composition variables pre policy, i.e., the percentage of learners at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction as well as a square and cubic term in the percentage of Tagalog-speaking learners (see Table 2). The pre-period is SY2008-2009 to SY2011-2012. See Figure 7 for per period coefficient estimates.

As regards our proxy for the average class size in elementary school, i.e. for students in Grades 1 to 6, the Pupils-to-Teacher ratio (PTR),[11] we find no significant impact of the policy for our preferred IV specification reported in column 2 of Table 5. This suggests that neither the decline in enrollment nor the departure of teachers dominates in tilting class size towards significantly decreasing nor increasing.

4.2 Years of Completed Education and Grade Completion (Census Data)

4.2.1 Empirical Approach

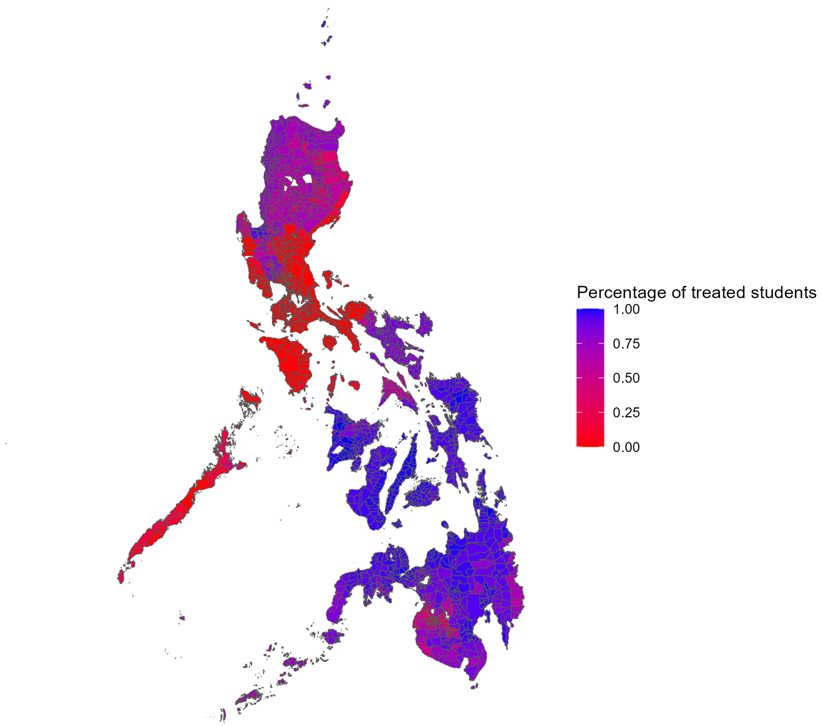
Combining data on the highest grade completed and a respondent's age in both the 2010 and 2020 Censuses of Population and Housing (CPH), we can use an identification strategy which builds upon the birth cohort difference-in-differences approach first used by Duflo (2001). Because we observe data for the same *age* cohorts both before, in the 2010 census round, and after implementation of the policy, in the 2020 census round, we can augment this approach by using a cohort triple difference (TD) design. In other words, we compare the differential years of completed education schedules across ages between treated and control municipalities both before and after the policy.

Note however, that we cannot directly match census respondents to schools in the DepEd database because respondents do not disclose their primary school in the census questionnaires. Instead, we match respondents' municipality of birth in 2020 (their municipality of residence in 2010[12]) with the municipality in which the schools are located. In order to create a measure of treatment intensity at the municipality level, Treat_m, which overcomes similar endogeneity concerns to those addressed by our IV approach, we exploit the first stage regression for which the coefficient estimates are presented in Table 2 and predict the probability of treatment for 34,807 public schools with information on the linguistic composition of the school in 2012-2013 (using the universe of Grades 4 to 6 students with mother tongue information). We then aggregate up the school-level predictions at the municipality level by weighting each school's predicted probability of treatment with the number of Grades 1 to 6 students in 2012-2013. This generates a municipality-level treatment intensity which

[11] For the PTR, in our regression analyses, we drop values below 12, the 1st percentile, to exclude nonsensical outliers below 1, as well as values above 96, the 99th percentile, to exclude the few outliers greater than 1000.

[12] Only the 2020 CPH contains information on respondents' municipality of birth.

Figure 4: Municipality-level Treatment Variation Across Space



Note: This figure shows the geospatial variation of our treatment intensity variable defined at the municipality level. Treatment intensity varies from 0 to 1 and is defined as the predicted percentage of treated students at the municipality level. It is constructed using the predicted values from the first stage regression presented in Table 2 for all schools with linguistic composition data. It is then aggregated up at the municipality level weighting each school's predicted probability of treatment with the size of the Grades 1 to 6 student population in 2012-2013. Darker shades of red represent a lower treatment intensity while darker shades of blue correspond to higher treatment intensities. White shading indicates municipalities excluded from the analysis (for which either census data or linguistic composition data is missing).

varies between 0 and 1 for the 1,627 municipalities in our sample, and corresponds to the predicted percentage of treated students.[13] Figure 4 shows the geospatial distribution of this variable across the Philippines. See Figure 6 in Appendix A.1 for the distribution of Treat$_m$ across census respondents aged 7 to 25.

We exploit variation in the exposure across ages, census rounds, and municipalities. In 2020, individuals aged 18 or older were beyond Grade 3 when the policy was implemented (i.e. they were 10 or older in 2012) and were thus not exposed to the policy as opposed to individuals aged 17 or younger in 2020. We estimate the

[13] Because this is a linear probability model, predictions are not bounded by 0 or 1 so we recode Treat$_m$ to be equal to 1 for the 17 municipalities with values between 1 and 1.025, and to be equal to 0 for the 5 municipalities with values between -0.009 and 0.

following regression equation:

$$
\begin{aligned}
Y_{iampr} = {} & \eta_{ar} + \eta_{mr} + \eta_{am} + \eta_{apr} \\
& + \beta_{\text{TD}}\, \mathbf{1}\{\text{Age}_a < 18\} \times \text{Treat}_m \times \mathbf{1}\{\text{Census}_r = 2020\} + \varepsilon_{iampr},
\end{aligned}
\tag{2}
$$

where Y_{iampr} is an outcome of interest for respondent i from the age a cohort, born in municipality m, in province p, in census round r. η_{ar}, η_{mr}, and, η_{am} correspond to the fixed effects of all double interactions between a, m and r. Once again, the inclusion of province × age × census round fixed effects ensures that we focus exclusively on deviations from a province-level trend. Treat_m is the municipality-level treatment intensity variable corresponding the the percentage of treated students. We cluster standard errors at the municipality level.

Identification. As discussed by Olden and Møen (2022), for this approach to have a causal interpretation, we must assume a *relative* parallel trend assumption holds. In our setting, and with a binary interpretation, this requires that the relative outcome in 2020 of those born in treated municipalities vs. control municipalities to trend (across birth cohorts) in the same way as the relative outcome in 2010 of those born in treated municipalities vs. control municipalities in the absence of treatment. Recall however that Treat_m varies continuously from 0 to 1. As a result, as Callaway et al. (2024) discuss for the DID setup, and in order for our estimates to have a causal interpretation, we must assume a stronger form of the parallel trend assumption: a *generalized parallel trends assumption* which involves potential outcomes under different doses of the treatment intensity. This assumes that the *observed* outcome changes for respondents in municipalities in each treatment intensity level reflect what would have happened—the counterfactual—for respondents in all other treatment intensity levels had they received that dose.

We also estimate the *dynamic* analog of equation (2) which allows for differential causal impacts across age cohorts relative to individuals aged 18.

$$
\begin{aligned}
Y_{iampr} = {} & \eta_{ar} + \eta_{mr} + \eta_{am} + \eta_{apr} \\
& + \sum_{\substack{h=7 \\ h \neq 18}}^{h=25} \tau_h \left(\mathbf{1}\{\text{Age}_a = h\} \times \text{Treat}_m \times \mathbf{1}\{\text{Census}_r = 2020\} \right) + \varepsilon_{iampr},
\end{aligned}
\tag{3}
$$

where Y_{iampr} is an outcome of interest for respondent i from the age a cohort, born in municipality m, in province p in census round r, η_{ar}, η_{mr}, η_{am} and η_{apr} are fixed effects. τ_h are the parameters of interest with the pre-trend hypothesis that $\tau_h = 0$ for $h \geq 19$, and causal effects of the program $\tau_h > 0$ for $h < 18$.

4.2.2 Triple-Difference Estimates on Highest Grade Completed

The results on enrollment raise the concern of potential compositional changes which could partly explain the policy impacts on test scores. If the best-performing students in treated schools left for control or private schools, this would bias treatment effects on test scores in a negative direction. To address this, we conduct analyses using Census data that are largely immune from such selection bias.

Our analyses here exploit respondents' municipality of birth in the 2020 census round. Using the birth locality for treatment assignment provides the benefit of obtaining coefficient estimates for the causal impact of the policy net of departures from treated schools, and thus more immune to concerns about compositional changes. The use of census data in 2020 also enables us to trace the impacts of this early educational quality shock on later educational attainment.

Table 6: Triple Difference: Impacts on Highest Grade Completed

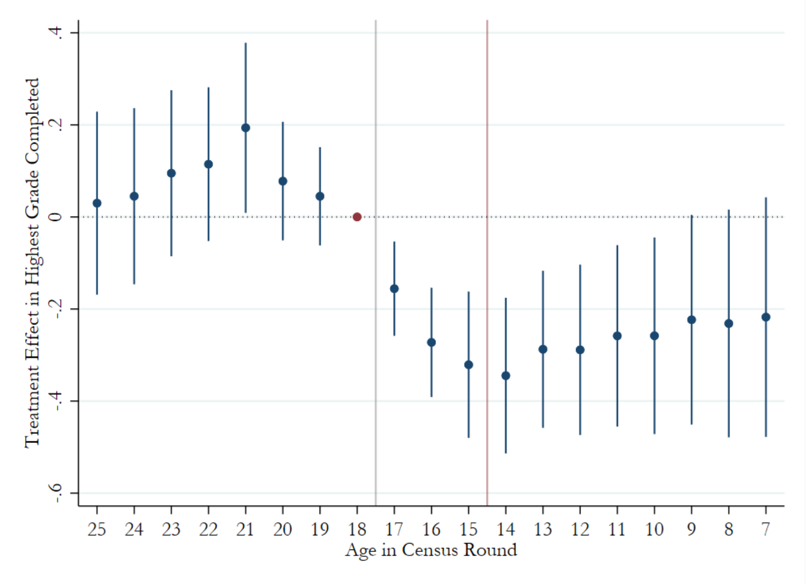
	Full sample		2010 Census	2020 Census
	TD		DID	DID
	(1)		(2)	(3)
Treat × (Age = 7-17) × (Census = 2020)	-0.334***	Treat × (Age = 7-17)	0.0442	-0.353**
	(0.079)		(0.204)	(0.177)
Control Mean (Age = 7-17), Pre	5.33	Control Mean (Age = 7-17)	5.33	5.84
Census × Age FE	Y	Age FE	Y	Y
Municipality × Census FE	Y	Municipality FE	Y	Y
Municipality × Age FE	Y			
Province × Age × Census FE	Y	Province × Age FE	Y	Y
Observations	73,267,484	Observations	35,025,793	38,241,691
Clusters (Municipalities)	1,627	Clusters (Municipalities)	1,585	1,627

Note: Standard errors are clustered at the municipality level. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the coefficient estimates on the highest grade completed (coded continuously in years from 0 to 16) for the triple interaction term from the estimation of equation (2) in column (1) and from the simple DID analogs for each census round in columns (2) and (3). Treatment intensity at the municipality level is defined as the percentage of treated students and varies continuously from 0 to 1. For the computation of the "control" mean among those aged 7 to 17 in the 2010 census round (pre policy), we use respondents from municipalities with treatment intensity less than or equal to 10%. The sample includes respondents aged 7 to 25 from the 2010 and 2020 censuses. Treated cohorts are those aged 7 to 17 in the 2020 census round (with varying levels of treatment intensity based on their municipality of birth; see Figure 4).

Table 6 shows coefficient estimates on the triple interaction term from the estimation of equation (2) on highest grade completed. Coefficient estimates for the simple cohort DID are also shown (with the 2010 Census DID corresponding to a placebo test). Figure 5 reports dynamic coefficient estimates (across age cohorts) from estimation of equation (3). We find a negative and statistically significant impact of the policy on the number of completed years of education. After controlling for province-level cohort trends, a respondent born in a municipality with 100% of treated students and exposed to the policy has, on average, 0.33 fewer years of completed education relative to a respondent born in a municipality with no treated students. Breaking down the triple difference into two double differences, column (2) in Table 6 shows a coefficient estimate close to zero and statistically indistiguishable from zero for the cohort DID placebo check using 2010 data only, and column (3) shows a cohort DID coefficient estimate using 2020 data only, exactly in line with the triple-difference estimate (with the main difference in statistical precision stemming from the inclusion of municipality × age FEs in the triple-difference estimation).

Figure 5 shows the estimated causal effects of the policy for each individual age cohort relative to the omitted 18 years old cohort (at the time of the census round). Recall that in 2020, students who were aged 8 or 9 and in Grade 3 in school year 2012-2013 are aged 16 or 17 at the time of their response in 2020. As a result, exposure varies non-linearly with age in 2020. Those aged 17 have up to 1 year of exposure (in Grade 3), those aged 16 have up to 2 years of exposure (in Grades 2 and 3), those aged 15 have up to 3 years of exposure (in Grades 1, 2 and 3), while those aged 10 to 14 have up to 4 years of exposure (from Kindergarten to Grade 3). Then, exposure decreases with age from age 9 to age 7 because students aged 7 or 8 have not yet reached Grade 3. There may also be a calendar time effect where younger cohorts who were exposed to the policy after additional years of implementation may have been less negatively impacted by the shock to education quality associated with the implementation challenges if schools developed solutions to improve teaching under the policy guidelines. The age profile of our dynamic triple-difference estimated coefficients are in line with these patterns. The magnitude of the impact increases from those aged 17 to those aged 14 in 2020, with those aged 14 the most negatively affected by the policy with a 0.34 decline in the number of completed years of education. The effect then appears to stabilize for younger cohorts at around -0.25 completed years of education.

Figure 5: Dynamic Impacts (across age cohorts) on Highest Grade Completed



Note: TD coefficient estimates (with 95% confidence intervals) from the estimation of equation (3). Respondents aged 18 at the time of the census round are the omitted age cohort. The dependent variable is a respondent's highest grade completed (coded continuously in years from 0 to 16). Treatment intensity at the municipality level is defined as the percentage of treated student and varies continuously from 0 to 1. The sample includes respondents aged 7 to 25 from the 2010 and 2020 censuses. Treated cohorts are those aged 7 to 17 in the 2020 census round (with varying levels of treatment intensity based on their municipality of birth; see Figure 4).

Using census data on the highest grade completed, we can also study the impacts of this policy change on the grade completion rates for individual grades from Grade 1 to Grade 10. Table A5 shows the coefficient estimates on the triple interaction term using Grade 1 to Grade 10 completion as the (binary) outcome in equation (2). Interestingly, we find that negative impacts in grade completion emerge in Grade 4, and increase slightly and stabilize in later grades up until Grade 8 with a statistically significant 1.7 percentage points decrease in grade completion.[14]

[14] As discussed above, because exposure varies across cohorts, the comparison of impacts across grades is not straightforward. The observed grade by grade pattern is consistent with Figure 5.

5 Conclusion

In this paper, we exploit a unique natural experiment in the Philippines to examine the long-term consequences of early education quality. We find that the quality of education in the first years of schooling has substantial and enduring effects on academic achievement and educational attainment.

The unexpected decline in educational quality resulting from the flawed implementation of a mother-tongue education policy allowed us to isolate the causal impact of early education quality on later outcomes. We found that students exposed to lower quality early education experienced significant declines in Grade 6 test scores across all subjects. Corresponding declines in student enrollment and teacher retention provide additional evidence of lower education quality in treated schools. In addition, analysis of census data reveals long-lasting impacts: fully-affected cohorts completed 0.3 fewer years of schooling by 2020, including for students whose last exposure to the policy was eight years in the past.

These results have important implications for both theory and policy. From a theoretical perspective, our findings support the hypothesis that early educational experiences play a crucial role in shaping long-term academic trajectories. The persistence of effects we observe underscores the complementarity between early and later human capital investments, as posited by Cunha and Heckman (2007). Our results also contribute to the ongoing debate about fade-out versus persistence of early education effects, providing evidence for the latter in the context of a broad, system-wide change in educational quality.

From a policy standpoint, our study highlights the critical importance of maintaining and improving the quality of early education. The substantial long-term costs associated with even temporary declines in educational quality suggest that policymakers should exercise extreme caution when implementing reforms that could potentially disrupt early learning environments. Furthermore, our findings emphasize the need for careful planning and piloting of educational reforms, particularly in multilingual contexts where language of instruction policies can have far-reaching consequences.

While our study focuses on the Philippines, the implications of our findings likely extend to other developing countries grappling with similar challenges in education policy and implementation. The magnitude and persistence of the effects we observe underscore the high stakes involved in early education quality and the potential for both significant gains from improvements and substantial losses from degradations in quality.

Future research could build on our findings by exploring the specific mechanisms through which early education quality affects long-term outcomes, and by investigating potential interventions to mitigate the negative impacts of temporary declines in educational quality. Additionally, longer-term follow-up studies could examine whether the effects we observe persist into adulthood, affecting labor market outcomes and other life circumstances.

In conclusion, our study provides robust causal evidence on the enduring impact of early education quality on academic achievement and educational attainment. These findings underscore the critical importance of prioritizing and protecting the quality of early education as a key component of human capital development.

References

Almond, D. and J. Currie (2011): "Killing Me Softly: The Fetal Origins Hypothesis," *Journal of Economic Perspectives*, 25, 153–172.

Angrist, J., A. Chin, and R. Godoy (2008): "Is Spanish-Only Schooling Responsible for the Puerto Rican Language Gap?" *Journal of Development Economics*, 85, 105–128.

Angrist, J. and V. Lavy (1997): "The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco," *Journal of Labor Economics*, 15, S48–S76.

Angrist, N., M. Ainomugisha, S. P. Bathena, P. Bergman, C. Crossley, C. Cullen, T. Letsomo, M. Matsheng, R. M. Panti, S. Sabarwal, et al. (2023): "Building resilient education systems: Evidence from large-scale randomized trials in five countries," *NBER Working Paper No. 31208.*

Angrist, N. and R. Meager (2023): "Implementation matters: Generalizing treatment effects in education," *Available at SSRN 4487496.*

Argaw, B. (2016): "Quasi-Experimental Evidence on the Effects of Mother Tongue-Based Education on Reading Skills and Early Labour Market Outcomes," *ZEW-Centre for European Economic Research Discussion Paper.*

Bailey, D. H., G. J. Duncan, F. Cunha, B. R. Foorman, and D. S. Yeager (2020): "Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions," *Psychological Science in the Public Interest*, 21, 55–97.

Ball, J. (2010): "Enhancing learning of children from diverse language backgrounds: Mother tongue-based bilingual or multilingual education in early childhood and early primary school years," *UNESCO.*

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukherji, M. Shotland, and M. Walton (2016): "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India," *NBER Working Paper No. 22746.*

Barro, R. J. and J. W. Lee (2013): "A New Data Set of Educational Attainment in the World, 1950–2010," *Journal of Development Economics*, 104, 184–198.

Bold, T., M. Kimenyi, G. Mwabu, J. Sandefur, et al. (2018): "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 168, 1–20.

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting Event-Study Designs: Robust and Efficient Estimation," *The Review of Economic Studies*, rdae007.

Bühmann, D. and B. Trudell (2008): *Mother Tongue Matters: Local Language as a Key to Effective Learning*, Paris: UNESCO.

Callaway, B., A. Goodman-Bacon, and P. H. Sant'Anna (2024): "Difference-in-differences with a continuous treatment," *NBER Working Paper No. 32117*.

Callaway, B. and P. H. C. Sant'Anna (2021): "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 225, 200–230.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024): "On binscatter," *American Economic Review*, 114, 1488–1514.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, 126, 1593–1660.

Cunha, F. and J. Heckman (2007): "The Technology of Skill Formation," *American Economic Review*, 97, 31–47.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006): "Interpreting the Evidence on Life Cycle Skill Formation," *Handbook of the Economics of Education*, 1, 697–812.

Currie, J. and D. Almond (2011): "Human Capital Development Before Age Five," in *Handbook of Labor Economics*, Elsevier, vol. 4, 1315–1486.

De Chaisemartin, C. and X. D'Haultfoeuille (2023): "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey," *The Econometrics Journal*, 26, C1–C30.

DepEd (2009): "DepEd Order No. 74 s. 2009: Institutionalizing Mother Tongue-Based Multilingual Education (MLE)," *Republic of the Philippines.*

——— (2012): "DepEd Order No. 16 s. 2012: Guidelines on the Implementation of the Mother Tongue-Based Multilingual Education (MTB-MLE)," *Republic of the Philippines.*

Duflo, E. (2001): "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91, 795–813.

Ebernhard, D. M., G. F. Simons, and C. D. Fennig (2023): "Ethnologue: Languages of the World," http://www.ethnologue.com.

Evans, D. K. and F. Yuan (2019): "Equivalent years of schooling: A metric to communicate learning gains in concrete terms," *World Bank Policy Research Working Paper.*

——— (2022): "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis*, 44, 532–540.

Glewwe, P. and K. Muralidharan (2016): "Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications," in *Handbook of the Economics of Education*, Elsevier, vol. 5, 653–743.

Hanson, M. (2024): "U.S. Public Education Spending Statistics," https://educationdata.org/public-education-spending-statistics, Accessed online on 07/14/2024.

Heckman, J. J. (2006): "Skill formation and the economics of investing in disadvantaged children," *Science*, 312, 1900–1902.

Heugh, K. (2012): "Theory and practice - language education models in Africa: research, design, decision-making and outcomes," in *Optimising Learning, Education and Publishing in Africa: The Language Factor: A Review and Analysis of Theory* *and Practice in Mother-Tongue and Bilingual Education in sub-Saharan Africa*, ed. by A. Ouane and C. Glanz, Hamburg: UNESCO, 105–156.

Kremer, M., C. Brannen, and R. Glennerster (2013): "The Challenge of Education and Learning in the Developing World," *Science*, 340, 297–300.

Laitin, D. D., R. Ramachandran, and S. L. Walter (2019): "The legacy of colonial language policies and their impact on student learning: Evidence from an experimental program in Cameroon," *Economic Development and Cultural Change*, 68, 239–272.

List, J. A. (2022): *The voltage effect: How to make good ideas great and great ideas scale*, Crown Currency.

Ludwig, J. and D. L. Miller (2007): "Does Head Start improve children's life chances? Evidence from a regression discontinuity design," *The Quarterly Journal of Economics*, 122, 159–208.

Mbiti, I., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani (2018): "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania," *The Quarterly Journal of Economics.*

Metila, R., L. Pradilla, and A. Williams (2016): "Investigating best practice in Mother Tongue-Based Multilingual Education (MTB-MLE) in the Philippines, Phase 2 progress report: Patterns of challenges and strategies in the implementation of mother tongue as medium of instruction in the early years: A nationwide study," Tech. rep., Assessment, Curriculum and Technology Research Centre (ACTRC).

Monje, J. D., A. C. J. Orbeta, K. A. Francisco-Abrigo, and E. M. Capones (2019): "Starting Where the Children Are: A Process Evaluation of the Mother Tongue-Based Multilingual Education Implementation," *PIDS Discussion Paper No. 2019-06.*

Mullis, I. V., M. O. Martin, P. Foy, D. L. Kelly, and B. Fishbein (2020): "TIMSS 2019 international results in mathematics and science," .

Muralidharan, K., A. Singh, and A. J. Ganimian (2019): "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, 109, 1426–60.

OECD (2018): "Programme for International Student Assessment (PISA) Results: Country Note Philippines," .

Olden, A. and J. Møen (2022): "The Triple Difference Estimator," *The Econometrics Journal*, 25, 531–553.

Piper, B. et al. (2018): "Examining the Secondary Effects of Mother-Tongue Literacy Instruction in Kenya: Impacts on Student Learning in English, Kiswahili, and Mathematics," *International Journal of Educational Development*, 59, 110–127.

Pritchett, L. (2013): *The Rebirth of Education: Schooling Ain't Learning*, CGD Books.

Pritchett, L., M. Woolcock, and M. Andrews (2013): "Looking like a state: techniques of persistent failure in state capability for implementation," *The Journal of Development Studies*, 49, 1–18.

Ramachandran, R. (2017): "Language Use in Education and Human Capital Formation: Evidence from the Ethiopian Educational Reform," *World Development*, 98, 195–213.

Rambachan, A. and J. Roth (2023): "A more credible approach to parallel trends," *Review of Economic Studies*, 90, 2555–2591.

Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores (2005): *Lifetime effects: The High/Scope Perry Preschool study through age 40*, Ypsilanti, MI: High/Scope Press.

Taylor, S. and M. von Fintel (2016): "Estimating the Impact of Language of Instruction in South African Primary Schools: A Fixed Effects Approach," *Economics of Education Review*, 50, 75–89.

Tupas, R. and I. P. Martin (2017): "Bilingual and Mother Tongue-Based Multilingual Education in the Philippines," in *Bilingual and Multilingual Education*, ed. by O. García, A. Lin, and S. May, Springer, Encyclopedia of Language and Education.

World Bank (2018): "Learning to Realize Education's Promise," *World Development Report. The World Bank*.

World Bank, UNESCO (2021): "Philippines - Learning Poverty Brief - 2021 (English)," Tech. rep.

Ye, T., A. Ertefaie, J. Flory, S. Hennessy, and D. S. Small (2023): "Instrumented difference-in-differences," *Biometrics*, 79, 569–581.

A Appendix

A.1 Additional variable definitions and statistics

A.1.1 Grade 3 & Grade 6 National Achievement Test (NAT) Scores

Standardization. Both Grade 3 and Grade 6 test scores were originally raw scores graded on arbitrary scales varying by subject (e.g. out of 10, out of 20, or even out of 27). Therefore, we standardize test scores, across test takers in our main sample of public schools with medium of instruction information, in each school year, and for each subject, using the mean and the standard deviation of test scores of students from *control* schools.

Results for Grade 3 should be interpreted with caution because of changes in the test content and test language resulting from the policy itself[15] and missing years with no or limited nationwide standardized testing (2014-2015 and 2015-2016; see Figure 3). Grade 6 test scores are a more attractive outcome to measure the impact of the policy because (i) they weren't affected by these changes, (ii) they were consistent across the study period, and (iii) they measure longer-term learning once students transitioned back to the dominant language for instruction.

A.1.2 Highest Grade Completed

The variable *Highest Grade Completed* measures the number of completed years of education and is recoded continuously from 0 to 16 using the homonymous categorical variable from the 2010 and 2020 census rounds (Philippine CPH). For responses ranging from *Grade 1* to *Grade 12*, the encoding is straigthforward. For responses from the 2010 census, before the shift to the K-12 basic education system in the Philippines[16] and when high school only went up to Grade 10, we encode the *1st Year of Post Secondary* or the *1st Year of College* as 11, following this logic up to the *6th Year of College or Higher* as 16. *Post Secondary Graduates* are assigned 12 completed years of education, while *Academic Degree Holders* 14. For responses from the 2020 census, we use the same encoding for responses corresponding to the old curriculum; and encode *Post-Secondary Undergraduates* and *Short-Cycle Tertiary Undergraduates* as 11, *Post-Secondary Non-tertiary Graduates* and *Short-cycle Tertiary Graduates* as

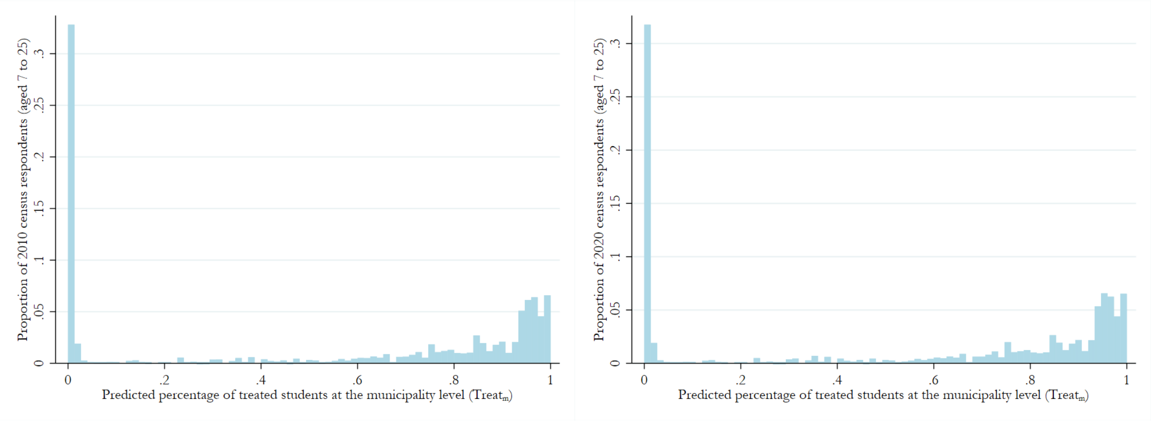
[15] The test changed from *NAT G3* from SY 2008-2009 to 2013-2014, to *LAPG G3* in 2015-2016 which did not include a mathematics test, to *ELLNA G3* in 2016-2017 and 2017-2018.

[16] The first cohort of Senior High School (SHS) students entered Grade 11 in SY 2016-2017.

12, *Bachelor's Degree Graduates* as 14, and *Master's Degree Graduates* and over as 16. This encoding reflects the fact that only older cohorts (in the 2020 census) who completed the K-10 basic education curriculum were old enough to reach advanced degrees.

A.1.3 Treatment Intensity at the Municipality Level

Figure 6: Distribution of census respondents across municipalities by treatment intensity



(a) 2010 Census **(b)** 2020 Census

Note: This figure shows histograms for the distribution of census respondents aged 7 to 25 (in the 2010 census on the left panel; in the 2020 census on the right panel) across values of treatment intensity Treat$_m$ defined at the municipality level for our analyses of census data. See section 4.2.1 for a description of the construction of this variable. The number of bins was set to 75. The mean value of Treat$_m$ across the 73,267,484 respondents used in our full sample (combining 2010 and 2020 respondents) is 54.1%, the 25th percentile is 1.0%, the median is 74.9%, the 75th percentile is 94.1%, and the standard deviation is 42.0%.

A.2 Additional Results

A.2.1 National Achievement Test Scores

Table A1: Impacts on Grade 3 test scores across subjects

	Overall		English		Filipino		Math.	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\bar{\tau}_{\text{post}}$	-0.003	0.023	0.008	-0.001	-0.011	-0.039	0.002	0.119
	(0.034)	(0.086)	(0.025)	(0.065)	(0.025)	(0.062)	(0.036)	(0.095)
Honest CI	[-0.162, 0.019]	[-0.378, 0.101]	[-0.143, 0.025]	[-0.375, 0.098]	[-0.158, 0.005]	[-0.443, 0.000]	[-0.164, 0.021]	[-0.309, 0.177]
Control Mean, Pre Period	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
School FE	Y	Y	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Obs. (Students)	856,610	856,610	1,012,160	1,012,160	1,012,160	1,012,160	856,610	856,610
Clusters (Schools)	23,712	23,712	23,808	23,808	23,808	23,808	23,712	23,712

Note: Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the coefficient estimates for the average causal effect across post-treatment periods from the estimation of equation (1) using Grade 3 test scores as the dependent variables. Overall test scores in Grade 3 are the average of English, Filipino and Mathematics test scores. Odd columns present estimates from the OLS specification while even columns show estimates from the instrumented DID specification where treatment status as the school level is instrumented with school-level linguistic composition variables pre policy, i.e., the percentage of learners at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction as well as a square and cubic term in the percentage of Tagalog-speaking learners (see Table 2). The pre-period is SY2008-2009 to SY2011-2012.

Table A2: Impacts on Grade 6 test scores across subjects

	Overall		English		Filipino		Math.		Science		Hekasi	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$\bar{\tau}_{\text{post}}$	-0.203***	-0.668***	-0.180***	-0.568***	-0.156***	-0.645***	-0.170***	-0.535***	-0.163***	-0.503***	-0.169***	-0.563***
	0.038	0.093	0.034	0.083	0.032	0.085	0.036	0.088	0.035	0.087	0.035	0.089
Honest CI	[-0.316, -0.147]	[-0.957, -0.532]	[-0.287, -0.135]	[-0.843, -0.458]	[-0.261, -0.123]	[-0.869, -0.505]	[-0.269, -0.111]	[-0.813, -0.409]	[-0.274, -0.117]	[-0.785, -0.380]	[-0.251, -0.095]	[-0.777, -0.372]
Control Mean, Pre Period	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs. (Students)	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528	1,102,528
Clusters (Schools)	23,395	23,395	23,395	23,395	23,395	23,395	23,395	23,395	23,395	23,395	23,395	23,395

Note: Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the coefficient estimates for the average causal effect across post-treatment periods from the estimation of equation (1) using Grade 6 test scores as the dependent variables. Overall test scores in Grade 6 are the average of English, Filipino, Mathematics, Science and Hekasi test scores. Odd columns present estimates from the OLS specification while even columns show estimates from the instrumented DID specification where treatment status as the school level is instrumented with school-level linguistic composition variables pre policy, i.e., the percentage of learners at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction as well as a square and cubic term in the percentage of Tagalog-speaking learners (see Table 2). The pre-period is SY2008-2009 to SY2014-2015.

A.2.2 Impacts on Enrollment Counts

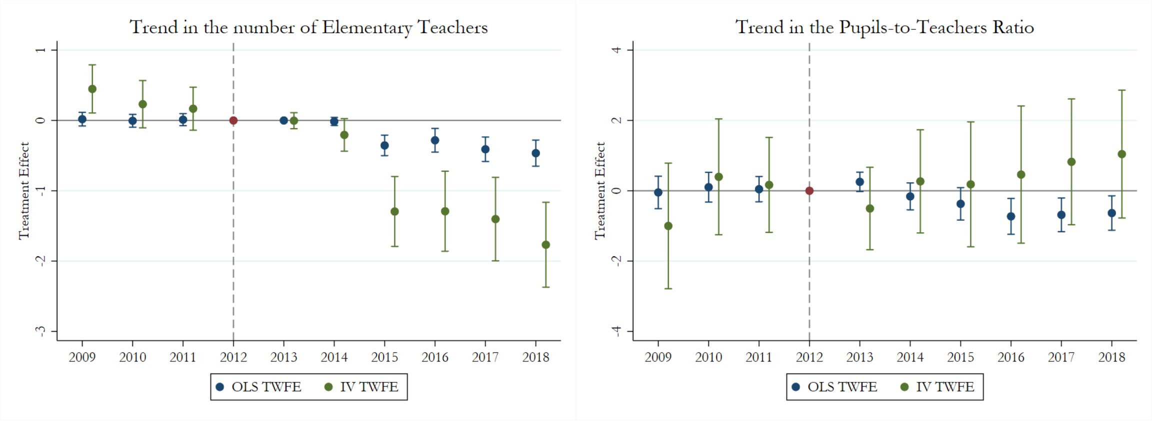
Table A3: OLS Estimates: Average Causal Effects on Grade 1 to Grade 6 Enrollment Counts

	Gr 1	Gr 2	Gr 3	Gr 4	Gr 5	Gr 6
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{\tau}_{\text{post}}$	3.373***	0.069	-0.583*	-2.119***	-1.501***	-1.646***
	(0.726)	(0.361)	(0.343)	(0.454)	(0.362)	(0.336)
Honest CI	[1.421, 6.401]	[-0.490, 2.434]	[-0.887, 1.758]	[-2.076, 0.084]	[-2.073, -0.357]	[-1.954, -0.557]
Control Mean, Pre Period	101.34	85.32	81.64	78.54	76.97	73.94
Year FE	Y	Y	Y	Y	Y	Y
School FE	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y
Observations	241,572	239,459	241,572	241,572	238,654	238,654
Clusters (Schools)	24,527	24,527	24,527	24,527	24,234	24,234

Note: Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.01.
This table shows the coefficient estimates on grade-level enrollment from Grade 1 to Grade 6 for the average causal effect across post-treatment periods from the estimation of equation (1) using the OLS specification. The pre-period is SY2008-2009 to SY2011-2012 for Gr 1 to 3. It is shifted by 1 year for Gr 4, two years for Gr 5, and three years for Gr 6 (allowing time for treated students to reach these grades).

A.2.3 Impacts on Teachers and the Pupils-to-Teachers Ratio

Figure 7: Dynamic Impacts on Teacher counts and the Pupils-to-Teacher ratio



(a) Teachers

(b) PTR

Note: Coefficient estimates (with 95% confidence intervals) from the estimation of equation (1) using the specification with school, school year, and province × year fixed effects. The dependent variable is elementary teacher counts in Panel (a), and the Pupils-to-Teacher Ratio (PTR), for which values below the 1st and above the 99th percentiles were dropped, in Panel (b). The pre-period is SY2008-2009 to SY2011-2012 while the post period is SY2012-2013 to SY2017-2018. IV estimates correspond to the instrumented DID specification where treatment status as the school level is instrumented with school-level linguistic composition variables pre policy, i.e., the percentage of learners at the school level whose mother tongue corresponds to each of the 19 languages offered as media of instruction. Standard errors are clustered at the school level.

Table A4: OLS estimates: Average Causal Effects on Teacher counts and the Pupils-to-Teacher ratio

	Teachers	PTR
	(1)	(2)
$\bar{\tau}_{\text{post}}$	-0.255***	-0.388**
	(0.057)	(0.172)
Honest CI (*smoothness restrictions*, $\bar{M} = 0$)	[-0.396, -0.063]	[-1.164, 0.279]
Control Mean, Pre Period	13.26	35.58
Year FE	Y	Y
School FE	Y	Y
Province × Year FE	Y	Y
Observations	239,140	236,494
Clusters (Schools)	24,461	24,431

Note: Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.01.

This table shows the coefficient estimates on elementary teacher count in column (1) and the Pupils-to-Teacher Ratio (PTR) trimmed at the 1st and 99th percentiles in column (2) for the average causal effect across post-treatment periods from the estimation of equation (1) using the OLS specification. The pre-period is SY2008-2009 to SY2011-2012. See Figure 7 for per period coefficient estimates.

A.2.4 Impacts on Grade Completion (Census)

Table A5: Triple Difference: Impacts on Grade Completion for each grade level

	Gr 1	Gr 2	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 9	Gr 10
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	β / SE	β / SE	β / SE	β / SE	β / SE	β / SE	β / SE	β / SE	β / SE	β / SE
Treat × (Age = X-17) × (Census = 2020)	0.00102	-0.00216	-0.00694	-0.0108*	-0.0122*	-0.0172**	-0.0196**	-0.0167**	-0.0124	-0.0130
	(0.003)	(0.004)	(0.005)	(0.006)	(0.007)	(0.007)	(0.008)	(0.007)	(0.008)	(0.009)
Control Mean (Age = X-17), Pre	0.97	0.95	0.93	0.92	0.90	0.87	0.81	0.76	0.68	0.60
Youngest cohort, X=	7	8	9	10	11	12	13	14	15	16
Census × Age FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Municipality × Census FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Municipality × Age FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Province × Age × Census FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	73,267,484	69,071,039	64,955,228	60,652,106	56,395,089	52,321,490	48,133,621	44,119,307	40,114,400	36,123,292
Clusters (Municipalities)	1,627	1,627	1,627	1,627	1,627	1,627	1,627	1,627	1,627	1,627

Note: Standard errors are clustered at the municipality level. *** p<0.01, ** p<0.05, * p<0.01.

This table shows the coefficient estimates on grade completion (across grade levels) for the triple interaction term from the estimation of equation (2). Treatment intensity at the municipality level is defined as the percentage of treated student and varies continuously from 0 to 1. For the computation of the "control" mean among those aged 7 to 17 in the 2010 census round (pre policy), we use respondents from municipalities with treatment intensity less than or equal to 10%. The sample includes respondents aged 7 to 25 from the 2010 and 2020 censuses. Treated cohorts are those aged 7 to 17 in the 2020 census round (with varying levels of treatment intensity based on their municipality of birth; see Figure 4).