RISE
RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

RISE Working Paper 20/042
July 2020

Myths of Official Measurement: Auditing and Improving Administrative Data in Developing Countries

Abhijeet Singh

Abstract

A central, yet understudied, component of state capacity is the quality of administrative data, which is a key enabler of effective policy making. I study this in the context of education systems in India. Using direct audit evidence, I show that levels of student achievement in Madhya Pradesh, from a large census covering approximately 7 million students annually, are severely inflated due to cheating. This distortion affects students at all levels of achievement but is particularly severe for low-performing students. In a follow-up randomized experiment, covering over 2400 schools in a different state (Andhra Pradesh), I evaluate whether tablet-based testing, which makes cheating harder, could reduce distortion. I find paper-based assessments proctored by teachers severely exaggerate achievement, in both private and government schools, but find no evidence of such distortion in tablet-based assessments. These results suggest that business-as-usual learning assessments may be compromised even without high-powered incentives, limiting their usefulness for policy action or research, although it may be possible to curtail such manipulation even at scale. This challenge of data corruptibility, and potentially evidence on reforms to address it, is likely to generalize across sectors.

UKaid from the British people
Australian Aid
Bill & Melinda Gates foundation
Oxford Policy Management
Blavatnik School of Government
University of Oxford

Myths of Official Measurement: Auditing and Improving Administrative Data in Developing Countries

Abhijeet Singh
Stockholm School of Economics

Acknowledgements:

I especially thank Karthik Muralidharan for his feedback and support at multiple stages of this project. I am grateful to Erich Battistin, Martina Bjorkman Nyqvist, Konrad Burchardi, Luis Crouch, Lee Crawfurd, Jonathan de Quidt, Jishnu Das, Tore Ellingsen, Clement Imbert, Gaurav Khanna, Derek Neal, Lant Pritchett, Mauricio Romero and several seminar participants for insightful comments. This project was supported by the ESRC Raising Learning Outcomes Initiative and Research in Improving Systems of Education (RISE) program funded by DFID. It would not have been possible without the support from the Governments of Madhya Pradesh and Andhra Pradesh – in particular, Ms. Deepti Gaur Mukherjee, Mr. Lokesh Jatav, Mr. K.P.S. Tomar, Ms. Sandhya Rani and Mr. Santhosh Singh – for which I am grateful. I am also grateful to staff at the Central Square Foundation, especially Rahul Ahluwalia, Saloni Gupta, Neil Maheshwari and Devika Kapadia, for their collaboration on program design and implementation for tablet-based testing in AP. Ramamurthy Sripada, Urmi Bhattacharya, Ghazal Gulati, Nawar Al Ebadi, Aditi Bhowmick, Sabareesh Ramachandran and Akankshita Dey provided outstanding field management and research assistance.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Department for International Development (DFID), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Singh, A. 2020. Test Scores and Educational Opportunities: Panel Evidence from Five Developing Countries. RISE Working Paper Series. 20/042. https://doi.org/10.35489/BSG-RISE-WP_2020/042

1 Introduction

State capacity for policy implementation depends on the availability of credible and comprehensive administrative data (Scott, 1998). Governments rely on such data, for example, to design and target interventions, provide incentives, and exert regulatory functions in many sectors, including education, health and social security. However, especially in settings of weak governance, underlying official data may themselves be severely compromised due to manipulation or corruption: Reinikka and Svensson (2004), for example, showed that budget data are unreliable for understanding local service provision or planning policy action in such settings. Yet, despite the constraint to effective service delivery that the distortion of micro-level data may pose, there is limited evidence from non-OECD settings on both the scale of the problem in large representative samples or of the effectiveness of interventions to improve them.

In this paper, I address both these gaps in the context of the education systems in low-and-middle-income countries (LMICs), focusing specifically on the reliability of administrative censuses on student achievement: these provide the primary measure of outcomes for compulsory schooling, a ubiquitous demonstration of frontline service delivery by governments. Such assessments also form the cornerstone of education reform globally to remedy a 'learning crisis' in LMICs, directly informing interventions and serving as policy targets.[1]My aim in the paper is two-fold. First, to measure the magnitude of distortion in administrative data using representative samples from two large Indian states. I estimate this directly by comparing officially reported achievement levels with results from independently proctored and graded retests. Second, I present causal estimates of the effectiveness of scalable policies to reduce such distortion and improve administrative data integrity.

The first part of this paper focuses on an annual standardized test of student achievement in the state of Madhya Pradesh (M.P.), which had a population of 72.6 million in 2011, that is administered to *all* students from Grades 1-8 in public schools (∼7 million students annually). This test, called *Pratibha Parv*, is used as a diagnostic exercise to focus on student achievement in the state education system but is not linked to student grade progression, teacher salaries or promotion, or similar high-stakes incentives. Designated as a "national best practice" in education (NITI Aayog, 2016), it is an exemplar of large-scale national assessments championed by international organizations and governments. In India, a similar census of student achievement is now proposed in all schools in Grades 3, 5 and 8 (Government of India, 2019).

[1]The 'learning crisis' refers to the widely-recognized problem of abysmally-poor learning outcomes for students in many developing countries even after several years of schooling. The very first policy response suggested by the World Bank in its flagship publication, the World Development Report, in 2018 was to assess learning "using well-designed student assessments to gauge the health of education systems [...] and using the resulting learning measures to spotlight hidden exclusions, make choices and evaluate progress"(World Bank, 2018). Proficiency-based measures are also central to national and global policy goals including the United Nations Sustainable Development Goals (Goal 4), adopted by 193 member-states.

The core of my analysis in M.P. directly compares students' reported responses in the official test in January 2017 with their responses to the *same* test questions on a retest, which was independently proctored and graded and serves as an external benchmark, in 283 primary and middle schools the following month. I use this to document three core results.

First, reported achievement levels are substantially overstated in the official data: the proportion of correct responses to the same multiple-choice questions is, on average, 38.9 percentage points (pp) higher in Math and 33.8 pp higher in Hindi in the official test, from a base of 25.1% and 37.9% correct responses in the retest in the two subjects respectively, i.e. a doubling of reported achievement. This leads the official assessments to be substantially misleading for inferences about average achievement: official data indicate that only ∼8% students fail in tests aligned to grade-level curricula while external assessments, such as the ASER reports, indicate that fewer than one-third of Grade V students can read a Grade II level text (Pratham, 2017).[2]

Second, this distortion – defined as the difference between the proportion of correct responses as reported in the official data and the retest – is present across the full distribution of prior achievement (as measured the previous school year) but is substantially higher for weaker students. However, even so, the distorted metric retains useful information about the ordinal ranking of schools and students. Both student achievement and the percentile ranking of schools exhibit strong positive correlation across years, implying substantial rank preservation even though levels are distorted. Thus, while the data appear unreliable for monitoring proficiency-based targets, they may yet be usable for other purposes.[3]

Third, this distortion is not merely an artefact of, say, a difference in student effort across the official test and the retest or fade-out of short-term test preparation between assessments. Rather, it likely reflects a combination of student copying and teacher-assisted manipulation of scores. In middle schools, the magnitude of this distortion is reduced by half in Mathematics to three-quarters in Hindi in classrooms where the test design mandated multiple booklets and external grading, both of which inhibit cheating, compared to other classes in the same school. Direct observation of testing in some schools and qualitative interviews with teachers and education officials further confirm cheating in this setting.

I follow up on these results by validating the generality of such distortion in a different state and experimentally evaluating a scalable policy intervention to reduce distortion. This second study covered Grade 4 students in over 2400 schools in one district in Andhra Pradesh state, who were randomly allocated to one of two treatment arms, and tested using standardized assessments

[2]Distortion seems also to have increased sharply over time: whereas in 2012, almost a quarter of the 112,000 schools fell in the "failing" D and E grades, this number had reduced to under 5% of schools by 2016.

[3]This preservation of ordinal information could plausibly reflect the fact that all administrative action based on this metric related to absolute achievement, not ordinal ranks. The use of ordinal information for policy purposes may distort this information in future implementation. I discuss these, and other, important caveats for the use of such metrics in Section 4.

in three subjects. In the benchmark case, covering 768 schools, school staff administered paper-based tests to students; these assessments used multiple test booklets but were graded centrally, to replicate the "best-case" scenario from M.P. In the intervention arm, covering 1694 schools, students took the same tests but now administered on tablets, which were brought to the school by a cluster resource person: as I discuss in Section 3, this is likely to make cheating harder for both students and teachers through multiple channels. Both paper and tablet assessments were conducted by government officials, as they would be in any census-based assessment, and not external staff. Thus, this study represents an evaluation at scale likely to identify a policy-relevant treatment effect (see Muralidharan and Niehaus, 2017, Vivalt, forthcoming).

I measure cheating in two ways. First, as in M.P., an externally-proctored paper-based retest audit, conducted in a subset of 117 schools, provides a direct measure of cheating. In this setting, it also helps to account for any differences in response patterns that arise solely from lower familiarity with tablet-based tests. Second, I use data on students' responses to individual test questions in the initial assessments to flag potential manipulation using the procedure implemented by Angrist, Battistin and Vuri (2017) in Italy.[4]This allows me to use data on all schools and also to compare the magnitude of cheating with the international literature, where retest audits of the type in this paper have not hitherto been possible and cheating has been detected through similar indirect procedures. I report two main results from this study.

First, there is evidence of substantial distortion in business-as-usual tests also in this setting. Using the Angrist, Battistin and Vuri (2017) procedure, 38-43% of classrooms in the paper-based testing arm are flagged for cheating. In the retest audit, students score 16-20 percentage points higher, on average, in the teacher-administered tests than in the independently-proctored retest. The magnitude of this distortion is comparable to the analogous sample from M.P. in Grade 8 (with multiple booklets and external grading) and indicates the generality of concerns around distortion even in the best-case scenario for typical government assessments.

Second, I find that distortion is substantially reduced in tablet-based testing. In contrast to paper-based assessments, only 2-5% of tablet-based test classrooms are flagged as potentially cheating by the procedure outlined in Angrist, Battistin and Vuri (2017). The difference between the official assessment and the retest is much smaller in the tablet-test arm and not statistically significantly different from zero in most cases.[5] Distortion appears to be more pronounced in private schools and larger classrooms.

[4]This is a procedure, adapted from the official practice in Italy to detect cheating in INVALSI exams, which uses item-level data to flag classrooms with suspicious response patterns (high within-class similarity of item-specific responses, high mean achievement, low variance, and a low proportion of unanswered questions). The exact procedure is explained further in Section 3 and Appendix C.

[5]Students answer between 2-4 percent fewer questions correctly in the tablet assessments than the (paper-based) retest, suggesting a small negative effect due to lack of familiarity with tablet-based assessments.

The main contribution of this study is to present novel evidence, at scale, of substantial distortion in reported outcomes in large-scale administrative data and options to improve it. The most closely related study in this vein is Duflo, Greenstone, Pande and Ryan (2013) who show that administrative data on pollution emissions is substantially misreported, undermining existing regulation, and that the assignment of external auditors reduced this misreporting. Results in this paper are complementary in two respects. First, I demonstrate that misreporting is stark even in the absence of direct financial incentives, which highlights that the threat to administrative data integrity is much broader than the financial corruption focused on in past work.[6] Second, I focus on a very different application. To my knowledge, no previous study in LMICs measures and reduces misreporting in *outcome* measures of public service delivery in administrative data, even though these underpin many policies focused on incentives, accountability or information: for instance, *all* reforms focused on 'results-based financing' for service delivery require high-quality and reliable measurement of outcomes.[7]

These results are particularly relevant for a large literature that evaluates interventions to improve student learning in developing countries (see Glewwe and Muralidharan (2016)). Many of the most promising reforms in this literature take reliable test scores as a pre-requisite. A prime example is performance-based pay in education: while this has been shown to have positive effects in large-scale experimental studies in India, East Africa and China (Muralidharan and Sundararaman 2011, Mbiti et al. 2019, Gilligan et al. 2018, Leaver et al. 2020, Loyalka et al., 2019), in each of these experiments, the test data triggering payments was collected either directly by research teams or their NGO partners. Similar concerns arise with, for example, report-card interventions where experimental studies report large gains (Andrabi, Das and Khwaja, 2017; Afridi, Barooah and Somanathan, 2018), but where the tests used to create report cards were conducted independently by the research teams. My results suggest that the unreliability of official data may be a binding constraint for scaling up such interventions and potentially much more important for their effectiveness at scale than, say, the optimal design of payment schemes and report cards with which this literature has primarily been concerned.

This paper also relates directly to studies that have examined manipulation of test scores by teachers or cheating by students.[8] This literature has, to my knowledge, focused exclusively on

[6] See e.g. Reinikka and Svensson (2004), Olken (2007) and Niehaus and Sukhtankar (2013) who focus on financial corruption in the public sector, and Martinelli and Parker (2009) who show misreporting to access conditional cash transfers .

[7] The World Bank reports 'results-based financing' in a number of sectors including energy, education, healthcare, solid waste management, water and sanitation and urban transport, which it coordinates through the Global Partnership for Results-Based Approaches (https://www.gprba.org/). The magnitude of resources committed to these initiatives can be large – for instance, a single initiative in healthcare, the Health Results Innovation Trust Fund, reports investments of $477 million from UK and Norwegian foreign aid, linked to 2 billion dollars in foreign aid from IDA.

[8] See, for instance, Jacob and Levitt (2003); Angrist et al. (2017); Dee et al. (2019); Diamond and Persson (2016); Wikström and Wikström (2005); Hinnerich and Vlachos (2017) documenting teacher-induced distortion

North America and Europe, typically in high-stakes settings, and relied on indirectly inferring cheating based on suspicious response patterns or bunching; in contrast, I measure cheating in a new setting using direct audits and an experimental evaluation of potential reform. Perhaps the closest parallel to my results in this literature come from recent investigations on cheating in Italy (Angrist, Battistin and Vuri, 2017; Battistin, De Nadai and Vuri, 2017; Bertoni, Brunello and Rocco, 2013). However, the magnitude of cheating I document is three times larger than even the southern Italian provinces: Battistin et al. (2017) report 14% presumed manipulators in math and 11% in language in the southern provinces (Italy-wide average of 6.6% and 5.6% respectively); in contrast, similar classification flags nearly 38-43% of classrooms in Andhra Pradesh. This banality of cheating forces a shift in focus from trying to detect (relatively-rare) cheaters, such as in Jacob and Levitt (2003), to reducing manipulability in the system as a whole. An unfortunate further consequence is that, despite governments' willingness to invest in such data capacity, comprehensive administrative data also appears less promising as a major resource for researchers in these settings than it has proven elsewhere.[9]

Finally, on an optimistic note, this paper adds to a body of work that evaluates technology-led solutions for system-wide public sector reform in developing countries. Recent papers have, for instance, looked at technology-aided solutions in procurement (Lewis-Faupel et al., 2016), social security payments (Banerjee, Duflo, Imbert, Mathew and Pande, 2016; Muralidharan, Niehaus and Sukhtankar, 2016),Muralidharan et al. (2018) and voting (Fujiwara, 2015). In these applications, technology plays three crucial roles: (a) it improves the timeliness, standardization and detail with which information becomes available, allowing for timely verification, (b) it reduces the risk of manipulation by circumventing corrupt actors, and (c) it allows for scale-up with fidelity. Tablet-based assessments play the same role here: they circumvent potential grade manipulation and make copying harder; they make detailed item-level data available for analysing suspicious response patterns; and they may have the potential for universal scale-up.

The rest of this paper is structured as follows: Section 2 presents the audit study in Madhya Pradesh and provides estimates of the magnitude of distortion; Section 3 presents results from the experimental evaluation of tablet-based assessments in Andhra Pradesh; Section 4 discusses the implications of these results for research and policy uses; Section 5 concludes.

---

and documenting copying by students (Borcan et al., 2017; Martinelli et al., 2018) . Of these, only Martinelli et al. (2018) use data from a developing country (Mexico) albeit in a very atypical setting of a high-stakes incentive program in 88 Mexican high schools.

[9]See Figlio, Karbownik and Salvanes (2016) for a comprehensive account of recent uses of administrative data in education research in economics. Such data have been, until recently, unavailable in most low and middle income countries, although some countries in Latin America provide important exceptions. In Chile, for instance, administrative test score data have been used to study the long-term impacts of influential policies like school vouchers (Hsieh and Urquiola, 2006), linked with birth registry records to assess cognitive impacts of early life circumstances (Bharadwaj, Løken and Neilson, 2013), as instruments used in policy design (such as community report cards in Mizala and Urquiola (2013)) and as an essential input for measuring the effects of randomized trials in the longer run (Neilson, Allende and Gallego, 2019).

2 Test distortion in Madhya Pradesh

2.1 Context and Policy Design

The first study in this paper is based in state of Madhya Pradesh (M.P.), which was India's fifth most populous state in 2011 with a population of around 72.6 million (72% rural). It is one of India's more deprived states, with a lower literacy rate and a higher poverty rate than the national average, and the country's largest population of Scheduled Tribes.

The public education system in MP illustrates the core challenges facing elementary school education in India. Learning levels of students are low - in 2016, only 31% of Grade 5 students in government schools were able to read a text at Grade 2 level; only 15.3% of Grade 5 students can do division (Pratham, 2017). These schools are small in size and, with an increase in private schools, have shrinking enrolment: whereas about 18% of rural primary schools had an enrolment below 60 students in 2010, this figure was over 40% in 2016. Teacher absenteeism in these schools is high, with 2010 levels estimated at about 26%; in a striking contrast from most other major states, which saw a reduction in teacher absenteeism, these levels had significantly worsened from 2003 estimates by 8 percentage points (Muralidharan et al., 2017).

The Government of Madhya Pradesh was an early adopter of large-scale assessments and instituted a state-wide census of student achievement in 2011 called *Pratibha Parv* (literally, "Festival of Talent"). This assessment is administered to all students in Grades 1-8 of the public schooling system, covering more than 110,000 schools and nearly 7 million students every year. As per government circulars, the assessments were expected to achieve multiple objectives: to understand the levels of learning and track progress of the system; to signal commitment and priority of the government towards learning metrics; to set up remedial measures to improve academic achievements; and to sensitize teachers, students and parents towards educational achievements of students. These assessments have been designated a national "best practice" in education (NITI Aayog, 2016) and similar student-level censuses are now planned nationally (Government of India, 2019).

Importantly, the assessment does not formally have high-stakes attached to it. Rather, it is intended to be a diagnostic exercise to identify students and schools most at need of support.[10] In practice, this is done by classifying students and schools into five letter-grades (from "A"

[10]This setting, thus, contrasts with the high-stakes environments in most previous papers such as from explicit monetary incentives (e.g Martinelli et al. 2018) or where the test results had significant long-term consequences for students (such as Borcan et al. 2017; Dee et al. 2019; Diamond and Persson 2016). The exception is the INVALSI test in Italy, where also the "purposes of the evaluation are to inform the central government about the general performance of the school system, and to offer schools a standardised reference to self-assess their strengths and weaknesses [...] tests are not formally high-stakes, because the allocation of resources to schools, the salary of teachers and the school career of students do not explicitly depend on test outcomes. Even so, pressure to perform well in the tests has been high because of the widespread expectations that they might be used at some point to evaluate teachers and schools."(Bertoni et al., 2013)

to "E") based on their aggregate score. Students in "D" or "E" grades are supposed to receive remedial instruction while schools classified as "D" or "E" are supposed to be prioritized for school improvement programs (such as *Shaala Siddhi*, evaluated in Muralidharan and Singh (2019)). However, these assessments play no role in students being promoted to the next grade (which is automatic until Grade 9), nor are they formally linked to any performance bonuses or criteria for promotion for the teachers (which is determined by civil service rules).

The assessment is conducted in December-January on the same days across the entire state. Students are tested in multiple subjects: in this paper, we will focus our attention on only the Mathematics and Hindi (language) assessments at the student level and the overall school aggregate at the school level. The assessment is centrally designed and, since this is the middle of the school-year, the content covers the latter half of the syllabus of the previous grade and the first half of the syllabus of the current grade. In 2016-17, the year of our audit, the test additionally covered some foundational skills. Questions in the test are a mix of multiple-choice items and those admitting open-ended responses. The test is exclusively written from Grade 3 onwards whereas it also includes individually-administered oral questions in Grades 1 and 2. In this study, we will restrict our investigation to the written assessments in Grades 3-8.

Cheating in these exams could stem from three main sources: (a) students could copy from each other or textbooks during the exam, (b) teachers could assist students in answering the exam questions and (c) teachers could inflate the grades that students receive. The state government is cognizant of the possibility of such cheating and has taken several measures to reduce it: question papers are sent to schools in sealed envelopes to prevent test questions being leaked to students before the day of the test; external officials are supposed to visit each school and verify that test administration is not being corrupted;and, importantly for our later analysis, in some grades, multiple sets of question papers are used to deter student copying and the answer booklets are sent to other schools in the same district for grading instead of being graded in the same school (which is the default). Yet, cheating is widely acknowledged to be common.

2.2 Data

The core of our analysis relies on independent data collected in 10 districts (out of 51 districts in the state): five districts in the Bhopal region and officially-notified tribal blocks in the five districts of the Indore region (Figure A.1). The cumulative population of these districts was over 15 million at the time of the last Population Census in 2011. We drew a sample for the independent data collection includes 3 schools each in 100 academic clusters (Jan Shiksha Kendras) across the ten districts i.e. a total of 300 schools. Table 1 presents summary statistics

of the sample schools from administrative data, comparing them to all schools in the sampling frame and to all schools in the study districts.[11]

Administrative data

Aggregate data, at the school level, on the official test scores in the Pratibha Parv assessments is available from 2012 to 2016. In addition, student-wise scores in each subject are maintained in physical registers in schools, which we transcribed for all students in the subsample for the official assessments preceding each round of our independent tests. In 2016-17, for the first time, schools were instructed to also record question-wise scores in each subject. In combination with the question papers used in the assessment, this allowed us to set up a direct audit study which can compare a student's response to the same test question when administered in an independently-proctored assessment in comparison to the officially-reported response.

Independent student assessment data

In the 298 of the 300 sample schools, we conducted three rounds of student assessments in mathematics and Hindi in July 2016, in February 2017 and in February 2018. These assessments were independently-proctored by surveyors hired and trained by the research team. Grading for the independent exams was done centrally. In this paper, we will rely on the data from February 2017, which will be compared to the official assessments in January 2017. We can match item-level data across the two assessments for students in 283 schools.[12]

Our independently-administered assessments included a number of items that were taken directly from the Pratibha Parv test papers; in the audit, I will rely only on these items. I will further restrict our attention to only those questions which were multiple-choice in the official assessment (and therefore in our retest). While this is not restrictive for the topics covered in mathematics, in Hindi the test questions common between our independent assessment and the official test typically cluster in areas related to reading comprehension and grammar.[13]

2.3 Diagnosing test manipulation in Madhya Pradesh

Table 2 presents the aggregate test score distribution in Pratibha Parv over time, with a special emphasis on the proportion of schools and students reported to fall in grades A to E from 2012, the second year of the assessment, and 2016 which is the assessment we will audit. During

[11]The data were collected as part of a large scale evaluation of the Shaala Siddhi school improvement program, which targeted improving school management in nearly 25,000 schools in the state. Please see Muralidharan and Singh (2019) for a detailed explanation of the sampling approach.

[12]Registers with item-level data were not available in all schools.

[13]This restricts the scope of our comparison to exclude, for example, any manipulation in grading that may be present in questions asking students to write a short essay. Our tests also do not include items which tested for rote memorization. This, again, primarily affects the audit of Hindi exams since a large part of the assessment in some grades relies on recall of specific facts in textbook chapters.

this period *reported* mean achievement increased substantially (from 54.5 points to 69 points, on a scale from 1-100) with a sharp reduction in the dispersion in test scores (from a standard deviation of 18.5 points to 11.6 points). This increase comes largely from the bottom-end of schools and students: whereas in 2012, a quarter of schools were assigned to "failing" grades (D and E), this proportion had sharply reduced to under 5% of schools by 2016; at the student level, whereas 14% were assigned an "E" grade in 2012, this declined to 1% by 2016. If genuine, this would indicate a remarkable improvement in the levels of student achievement in the state with a near-elimination of very low performance by students. Unfortunately, this seems unlikely. While independent data also show some improvement over time, the magnitude of such improvement is substantially lower: in 2012, 7% of Grade 3 students in government schools could read a Grade 2 level text and 6.8% of students could do a simple subtraction problem; by 2016, these proportions were still only at 10.2% and 8.4% (Pratham, 2017).

The core of my analysis focuses on comparing item-level student responses, for the same students, across the 2016-17 Pratibha Parv assessments, administered from 18-20 January, and the independent student assessment in February 2017, to quantify the magnitude of distortion. Figure 1 presents the core result of this comparison, restricting focus to only multiple-choice questions.[14] Each dot in the plot represents an individual test question. The horizontal axis presents the proportion of correct responses as reported in the official test data; the vertical axis presents the proportion of correct responses in the retest. The sample of students used for generating these graphs is restricted to students who took both assessments.[15] With the exception of a single question in math, the proportion of correct responses is substantially higher for all questions in the official test than in the independent assessment. The line of best fit indicates that, for a mathematics item where 80% of students were reported to answer correctly in the administrative tests, only about 30% of them could answer it correctly also in the retest. The magnitude of this discrepancy seems to be higher in Mathematics than in Hindi.

Figure 2 attempts to understand whether the extent of this discrepancy varies across students at different levels of achievement. The vertical axis in the left panel shows the difference between the proportion of correct responses in the official test and the retest on the common items at the student level. This is bounded between 1 — which would indicate that the student was reported to answer all items correctly in the official test but did not answer a single one correctly in the retest — and -1 for the opposite case. I investigate non-parametrically whether the discrepancy varies by the percentile of the student in the overall score distribution in that subject/class in

[14] There were 69 such items, split across the two subjects.

[15] Note that not all students can be matched across the two assessments, mainly because either the student was absent on the day of the independent test or that we are not able to match students due to missing identifiers. Figures A.2 in the appendix compares the distribution of achievement on the official test, which is available for nearly all students, for the sample whom we can match across the two datasets to those we cannot: these distributions are near-identical providing confidence that the sample used in our analysis is representative of these schools.

the previous school year, which is plotted on the horizontal axis.[16] This discrepancy seems to be substantially higher for weaker students but is positive at all parts of the distribution. In Hindi, the discrepancy is just above 40 percentage points in the bottom decile while being under 15 percentage points in the top decile. The decline is less stark in math (with higher discrepancy): in the bottom decile, this is close to 45 percentage points while in the top decile, this is about 15 percentage points lower.

Despite substantial distortion in levels of achievement, there is clearly some useful variation in the administrative data. In Figure 1, the probability of correct responses for a given question is positively correlated across assessments in both subjects. In Figure 2, we see that there is correlation in students' performance the previous year and their scores in both our independent assessment and the official assessment in the following year. Finally, in Figure 3, I relate the absolute reported achievement at the school level, and the percentile rank in the within-year distribution, to the same variable the previous year using the non-parametric methods of Cattaneo et al. (2019). The left panel shows clearly the the "pulling-in" of low-achieving schools over time: in 2015-16 and 2016-17, in comparison to previous years, fewer schools are reported as attaining low marks and, those that do, have higher predicted marks the next year. Yet, scores are clearly persistent and, unlike absolute achievement especially at the bottom end, the persistence of school rankings appears to be very stable across successive years with no change in this relationship over this period. The retention and stability of ordinal information in the tests, despite widespread manipulation, is remarkable and has potential implications for appropriate uses of these data, a point to which I return in Section 4.

2.4 Is this mismatch due to manipulation?

Figures 1 and 2, although strongly suggestive of manipulation, are not conclusive. If students exert lower effort in the independent assessments, perceiving them to be lower-stakes than official tests administered by their teachers, that could in principle explain the discrepancy between the two assessments. Further, the official tests are scheduled and pre-announced, whereas the independent assessments are not. If anticipated exams are preceded by extensive coaching and revision, as is common in many education systems including in India, the discrepancy could

[16] Plotting the two tests in 2017 against the test scores from the previous academic year (administered in December 2015), instead of against each other, reduces the threat to inference posed by mean reversion, which may otherwise be substantial (Chay et al., 2005; Jacob and Rothstein, 2016). Specifically, as the right-panel in Figure 2 shows, the distortion at different parts of the achievement distribution will only be confounded if there is *differential* mean reversion, from Dec 2015, between tests administered in Jan 2017 and Feb 2017.

One drawback of Figure 2 is that , by necessity, it is restricted to only those students for whom school records could be transcribed and matched across years. This leads in particular to dropping all students in Grade 6 (which is the fresh intake in middle schools) and any students who had transferred to the sample school in only the current year.

merely reflect this "test preparation effect". I take three distinct approaches to document that it is unlikely that these explanations account for the discrepancy.

The first approach utilizes the fact that, in 2016/17, the state government had taken additional steps to counter test score manipulation in some grades. In particular, the assessment introduced multiple test booklets in Grades 3, 5 and 8 and also asked for Grade 8 answer scripts to be sent to a different school for grading.[17] I study whether, in the same schools, the difference between the official and retest data is lower in Grade 8 as compared to other grades in the same school.[18]If this discrepancy between performance in the two assessments is mostly caused by factors such as differences in student effort on the two tests, or the rapid fade-out of short-term test preparation, then we should see no difference in this discrepancy for Grade 8, where cheating was harder, as opposed to other classes in the same school. Table 3 reports the results of this investigation. Student-level average test scores, on the same multiple-choice questions, are higher by ~54 pp in Math in Grade 7 on the official assessment than the retest, but this discrepancy is reduced by ∼50% in Grade 8 (Column 1); conditioning on school fixed effects does not affect these estimates. In Hindi, the mismatch is lower at about 36 percentage points in Grades 6 and 7, but is reduced by 27 percentage points in Grade 8 (Columns 3 and 4). That the magnitude of the discrepancy is cut so sharply by multiple booklets and external grading suggests that a substantial part of this discrepancy is accounted for by cheating by students and teachers.

The second, and more direct approach, relies on direct observation of the official tests in classrooms. In 2016-17 and 2017-18, we conducted observations of the test administration in 52 classrooms across 17 schools in four districts. These observations were done with authorization from the state government but schools were not informed of our visits in advance.[19] In nearly all classrooms, we found some form of student copying: most teachers did not actively try to control such copying; in 40% of classrooms, the teacher left the classroom for at least part of the test administration, leaving students unsupervised. Teachers also directly contributed to cheating: it was common to see them provide "hints" to the correct answer to students and to help them erase and correct the answer, if the student had answered incorrectly. External monitors, while formally appointed for each school, are rarely present in classrooms during the

[17]Answer-scripts for Grade 5 were externally graded until 2015-16 but this was not required in 2016-17.

[18]Unfortunately, I cannot do this exercise also in primary schools (Grades 1-5). The only available "control" grade, with written assessments and which did not have multiple sets, is Grade 4 but there are very few multiple choice items in common between the Pratibha Parv and our retest. Comparing discrepancy across open-ended and multiple-choice items, even where test questions can be graded unambiguously (such as in math computation questions), is not valid because of the non-zero chance of answering correctly in both the original and later test through random guessing.

[19]The presence of the observers was, of course, known to school staff during the tests. Any "Hawthorne effects" arising from this should have reduced the incidence of cheating in the schools observed. The schools selected were outside our main evaluation sample and were purposively selected to cover multiple districts and types of schools. A detailed description of this exercise is provided in Appendix B.

assessment and, if they did come to the school, stayed only a short while and did not directly affect test administration.

The third approach relies on qualitative interviews with teachers and education department officials. These semi-structured open-ended interviews were conducted in 2018 and focused on a range of topics, including the nature of assessments in the state. Details of this exercise are provided in Appendix B, along with selected excerpts from the transcripts, which I summarize here briefly. Teachers and education department officials can, almost universally explain the purpose of the tests but differ in their assessments of its usefulness; some, at least, are dismissive about any usefulness at all. Some teachers also acknowledge cheating by students in the exam, acknowledge that their students cannot engage with the level of questions and admit assisting students with the exam. These interviews also highlight that, despite the lack of formal high-stakes incentives, at least some teachers fear the potential for consequences and do not want to be *seen* as underperforming. As summarized by one teacher:

> On paper, all achievement is very good, all students are in A Grade. [...] If we say that all of these students, whom we have shown to be A grade, are actually only at C grade level, then there will be someone here from the administration with a stick asking why this is the case. He will not listen that there were many other duties during the year that kept us away from school.
>
> [...] This is all only on paper and it is wrong. We send A grade results, the Jan Shikha Kendra compares our result and claims accolades that the cluster is doing so well, he sends it to the BRC, who sends it to district-level officials and then finally when the state-level officials look at this on the online portal, they think the school is functioning very well – it's only when they compare what they see on the portal to what they actually find when they come to the field that they have any understanding.[20]

This may help explain why teachers may grade students up or providing assistance during the test, although it is unlikely to be the only reason.[21]Taken together, the evidence from these three distinct approaches confirms that the substantially worse student performance in the independent tests, as compared to the official assessments, reflects cheating in the administrative data.

[20]The Jan Shikshak, also called the Cluster Resource Coordinator, and the Block Resource Coordinator (BRC) are the two levels of the education bureaucracy and supervision immediately above the school level.

[21]Teachers may, for instance, decide to actively assist to make up the socioeconomic disadvantage of their students, if they think that is the main reason for poor learning levels. This would be consistent with some qualitative reports (see Appendix B) and would also be similar to the pattern of teacher discretion in Linden and Shastry (2012). Alternatively, they may misreport because they associate higher achievement of their students with status: Martinelli and Parker (2009) document that households in Mexico sometimes over-report their assets mainly to avoid the loss of status in admitting that they do not own a particular consumer durable, even though such a report reduces their likelihood of getting a cash transfer.

3 Reducing test manipulation in Andhra Pradesh

Section 2 documents substantial manipulation in administrative data in Madhya Pradesh. In this section, I present results from a second state to establish the external validity of these concerns and to evaluate potential solutions for reducing such distortion at scale.

3.1 Background and Setting

The second study is based in the state of Andhra Pradesh, which is the seventh largest state in India by area and had a population of 49.3 million in 2011.[22]Within India, it is known for above-average state capacity in the implementation of government programs, especially in social sector initiatives.[23] In 2018, 59.7% of Grade V students, across private and government schools, could read a Grade II level text and 39.3% could divide (Pratham, 2019); these figures are better than the averages for Madhya Pradesh (which is a poorly-performing state within India) or for the whole country but they are still poor and worse than in 2012.

This experiment was carried out in Prakasam district (see Fig A.3), which is close to state-level averages in various educational indicators: roughly 35.5% of children between 6-14 years of age are enrolled in private schools (compared with 35.2% state-wide), 43.5% of students in Grades 3-5 can read a Grade 2 level text (41.5% state-wide), and 60% of students (56.6% state-wide) in these grades can do at least subtraction (Pratham, 2019).

3.2 Experiment Design

The Government of Andhra Pradesh, as part of a larger initiative of school transformation (Badi Parivartana), wanted to test and implement a system of providing detailed report cards to parents which provided information on achievement levels of their children but also of all schools in the geographical area. Report card interventions like these require information on all schools and thus remain particularly susceptible to cheating and test manipulation. As a potential solution to this central challenge, the government authorized a large-scale pilot evaluation of tablet-based assessments which may reduce student copying and teacher-aided manipulation and make item-level data easily available for further verification. It would also improve the timeliness of data availability, while avoiding the incremental burden of grading and data entry typically faced by teachers.

[22]The state has also been the setting for much of the research in economics of education in India, having been the setting of the AP Randomized Studies in Education (see e.g. (Muralidharan and Sundararaman, 2010, 2011, 2015)and the Young Lives study (see e.g. (Singh, 2015, 2019)).

[23]For example, Andhra Pradesh was designated as a "star state" for the implementation of the National Rural Employment Guarantee Scheme, as evaluated by Imbert and Papp (2015) and was also the setting for the improved implementation of the program in later years through biometric identification of beneficiaries Muralidharan et al. (2016).

The effectiveness of such measures is not known in developing countries. Moreover, because any system of tablet-based assessments is likely to involve more complex logistics than a (long-running) paper-based system of assessments, it is unclear that any such models would be implementable by government systems at sufficient scale, which was an important concern for the Government of Andhra Pradesh.[24]

To study this, we set up a large-scale experiment covering *all schools* with at least 5 students enrolled in Grade 4 in one district in February 2019. Randomization was carried out not at the level of the individual school but at the level of an academic cluster which typically covers multiple villages. This was to keep the mode of testing unchanged within a single educational market so that schools within the same community could be fairly compared in student report cards. Out of 284 clusters, 196 were assigned for tablet-based testing and the remainder 88 to paper-based testing. All schools within a cluster were assigned to be tested using the same protocol (whether paper or tablets).[25] In the final sample, 768 schools were assigned for paper-based testing and 1694 to tablet-based testing. Table 4 presents descriptive statistics based on administrative data on the tablet and paper based assessment schools.

Students, across both treatment arms, were assessed in three subjects – Mathematics, Telugu (the official state language), and English – using the same test papers in February 2019. The tests were designed centrally and intended to capture a wide range of variation. All questions in the test booklets were multiple choice items. In each subject, to deter student cheating, three test booklets ("sets") were created which had 80% of items in common but with some distinct items and with a distinct ordering of test questions. In schools assigned to paper-based assessment, the test booklets were sent with clear instructions for schools to administer the tests and for the answer scripts to be returned to the Mandal Education Office (the administrative unit above schools). The reference mode of testing thus closely resembles the best-case scenario in paper-based testing in Madhya Pradesh, with multiple choice items, multiple sets of question papers and external grading.[26] In the tablet testing arm, the tests were carried out in a staggered

[24]The use of computer-based assessments is, of course, ubiquitous in other settings including in e.g. the SAT and GRE tests administered in many countries by ETS. However, the magnitude of these particular challenges is likely to be much more severe in public education systems in developing country settings than the typical OECD or elite private school settings where such testing is common.

[25]There was one pre-determined deviation from the initially-assigned status. Academic clusters, for the most part, nest villages/urban wards but not always. In such cases, if cluster A and B were in different treatment arms, the treatment status of all schools in the village was reassigned to be the same. For instance, if a village had 6 schools, 4 of which are in cluster A and 2 in cluster B, the treatment status of all schools in the village was reassigned to be the same as Cluster A. Such reassignment affected 191 schools out of 2462 schools, which are disproportionately urban areas.

[26]In contrast, for example, the larger standardized end-of-year test administered by the Government of Andhra Pradesh in all schools (Summative Assessment 2), which is typically used as the basis of achievement comparison in the state and with results made available online, is graded by the teachers in the same school and only the grades themselves are transmitted. Thus estimates of cheating in this experiment may be under-estimates of the true prevalence of manipulation in state administrative exams.

manner across schools. Tablets were taken to the school by a Cluster Resource Person (CRP), each student was given an individual tablet to work on and the test booklet for each subject was decided by the software directly.

In comparison to paper-based testing, tablet-based tests as administered here may reduce manipulation through several channels. First, they make student copying harder since students only see one question at a time and it is much harder to copy from the booklets of nearby students. Second, for the same reason, it is harder for teachers to help all students, or to see whether they have answered correctly (and provide the right answers). Teachers also cannot retrospectively erase and correct answers after the test has ended. Third, the mode of administration, since it required the CRP to take the tablets to the school and collect them after the test, ensured external observers were actually present at the time of testing (as opposed to the Pratibha Parv test in MP, where they were assigned to be present but rarely were). My goal here is not to distinguish between these potential channels but rather to evaluate whether this change in the mode of testing is (a) achievable at scale and implemented by government officials and teachers and (b) delivers test metrics with less manipulation.[27] Program implementation by government staff reduces the concerns of external validity which accompany high-fidelity implementation by motivated NGOs or research teams (Bold et al. 2018; Vivalt forthcoming). Thus the experiment is an "evaluation at scale" in all three respects highlighted by Muralidharan and Niehaus (2017) i.e. representative of large populations, studying implementation across a large number of treated units, and studying implementation at a large unit (to estimate effects that are net of spillovers).

The analysis in this study will compare the responses of students who were assessed using tablet to those who were assessed using paper booklets. A potential problem in this comparison is that students may underperform in tablet-based assessments, not only due to potential distortion in paper-based tests but also because they are not familiar with tablet-based tests. Thus, we randomly sampled 120 schools, spread equally across the paper and tablet-based testing arms, and retested students using a traditional paper assessment but with external proctoring and test administration by the research team. Table 4 presents sample characteristics for this set of schools. This retest was conducted within two weeks of the original assessment but could only be completed in 117 schools.[28]

[27]Put differently, I measure a composite policy effect which includes both the direct effects of the technology in inhibiting cheating and the indirect effect from complementing existing monitoring capacity in the system.

[28]This was due to ambiguity in finding and matching the relevant sampled school from the administrative dataset to the school on the ground (partly due to issues in school codes). Because there was a very tight window in which all retests needed to be completed before the closure of schools for the year, this resulted in our inability to track back 3 schools. Sample characteristics are balanced across arms in the restricted sample of 117 schools as well (Table A.1).

3.3 Results

3.3.1 Cheating in paper and tablet tests

Figure 4(a) presents the distribution of percentage correct in the official tests, separately for the paper-based and tablet-based assessments, for the full sample of students who were tested. Students who were tested on paper score much higher — by about 28 pp higher in mathematics, 26 pp in English and 21 pp in Telugu —than students who were tested on tablets. The resulting distribution of achievement in paper-based tests is very negatively-skewed, with substantial ceiling effects, whereas tablet-based tests provide a more bell-shaped distribution as would be typical in well-designed assessments.

Figure 4b shows the difference between the two testing arms, aggregated at the individual item level (rather than student level). As in Figure 1, each marker shows the proportion correctly answered for each individual question in the tablet and the paper testing arms. The proportion of students answering each item correctly is substantially higher in the paper-based tests than in the tablet-based assessments. This highlights that conclusions about student achievement may differ drastically if administering the assessment in one mode versus the other, i.e. the policy-relevant choice facing governments.

The pattern in Figure4 could be accounted for entirely by the lack of familiarity with tablet-based assessments and do not necessarily imply distortion in the paper-based assessments. To assess whether this difference is caused by cheating, I first use an indirect approach similar to Angrist et al. (2017) which uses the full item-level data from the paper- and tablet-based assessments and then a more direct comparison with an independently-proctored retest. This approach uses item-level data at the student level to generate four summary statistics at the classroom level: (a) the mean percentage correct, (b) the variance of percentage correct, (c) the proportion of non-missing answers and (d) an index of homogeneity of answer options in the classroom. These data are reduced to two principal components, following which schools are classified into clusters using a hard k-means clustering approach. Essentially, the method flags classrooms which have a very high mean score, low variance, low heterogeneity in answer options and low proportion of answers with missing responses.[29]I pool the item-level data for both modes of assessment and run this algorithm to flag classrooms with suspected manipulation separately by subject.

Figure 5 compares suspected cheating in the two modes of test administration. 38-43% the classrooms in the paper assessment are flagged in the extreme cluster using this approach in each subject. This is much higher even than the figures reported by Battistin et al. (2017) in southern Italy, where the proportion of suspected manipulators was assessed to be between 11-16%. In contrast, only 2-5% of the classrooms are similarly flagged in the tablet assessment.

[29]Please see Appendix C for a detailed description of the procedure and the resulting clusters.

In the subset of schools where we conducted a retest, I can compute a more direct measure of cheating. Figure 6(a) shows, for individual students matched across the tests, the difference in percentage correct across items common in the official tests and the retest – for the tablet based tests, the differences are centered close to zero (indicating little deviation on average) but for paper-based tests, scores are clearly higher in the official tests than the retest. Figure 6(b) plots the correspondence for individual test items: for all items, paper-based official tests significantly over-report student achievement whereas tablet tests seem to correspond very closely with the retest, although possibly doing a little worse (which would be consistent with small negative effect of the lack of familiarity with tablet based testing). This is confirmed in Table 5: paper tests exaggerate performance by 16-20 percentage points in each subject. I cannot reject equality between the tablet tests and the retest at the 5% level for any subject, although point estimates consistently point to a negative effect of 2-4 percentage points in the tablet-based assessment, which could be reflective of the lack of familiarity with tablet-based assessments.

The retest also allows me to directly validate the procedure used in Angrist et al. (2017). Specifically, for the students in the retest sample, I can compare our direct measure of cheating (the difference between the official test and the audit) across schools identified by the Angrist et al. procedure. This comparison is presented in Appendix figure A.4: there is little disagreement, on average, between the official test and the audit in schools which were not flagged by the Angrist et al procedure, whereas this difference is large (17-21 p.p.) in schools that were flagged.This suggests substantial agreement across the two ways of classifying manipulation.[30]

Finally, note that comparing the retest with the original test is likely to provide an underestimate of any reduction in distortion in moving from paper to tablet assessments. The official paper-based assessments, with multiple sets and external grading, correspond to the low-distortion case in Madhya Pradesh and already represents the best practice in the government sector. To the extent that most such assessments are typically graded within the same school (as also the case for most standardized tests in Andhra Pradesh), we would expect distortion under business-as-usual to be even higher. Second, it is possible that there are learning effects between the original assessment and the retest, which was conducted after a very short window. Any such effects are particularly likely to be concentrated in the paper-based assessment arm: it is possible that answering the same question in the same format is easier the second time around; it is also possible that, in the paper testing arm, teachers used the spare test booklets to provide

[30]This is not merely a difference between the tablet and paper test arms. Restricting the validation only to the paper-based treatment arm, where about 45% of schools in the retest sample are classified as cheating by the Angrist et al procedure, shows similar results: the disagreement between the reported scores and the audit is substantially higher in schools that had been flagged as cheating.

revision to students.[31]Evidence that suggests that this is, in fact, the case since students who were initially tested on paper perform somewhat better in the retest (Appendix Figure A.5).

3.3.2 Correlates of cheating

The large sample of schools in the A.P. experiment also allows for an opportunity to assess if the degree of cheating differs by observed characteristics of the classroom. I investigate this by testing whether the difference in percentage correct between the tablet and paper-based assessments, in the full sample of schools, differs by observable characteristics of schools and classrooms.

I look at heterogeneity along four dimensions – whether the school is private, the number of students enrolled in Grade 4 (which was the grade tested), the proportion of students who are girls, and the number of inspections in the school by cluster or block-level resource coordinators as per the official data. Unfortunately, I have limited information available about the students themselves and cannot test for heterogeneity on individual characteristics other than gender.

Results are presented in Table 6. Across all three subjects, it appears that the difference in performance in the paper and tablet tests is greater in private schools than in government schools. The magnitude of this difference is meaningful: whereas government schools students perform between 18-24 percentage points worse on tablets than on paper, this difference is larger by 3-8 percentage points in private schools. It appears also that cheating is greater in larger classrooms although the magnitude of this relationship is much more modest: an increase in the number of students by 10 students only increases cheating by $\sim$0.8 percentage points on average. The difference between tablet and paper assessments also appears to be decreasing in the proportion of girls in the classroom: a classroom with only boys is predicted to have 32 percentage points lower scores on tablets, which is reduced by 11-12 percentage points in a girls-only classroom in Math and Telugu although the coefficient is both small and statistically insignificant in English. Finally, there is very little evidence of the number of inspections, which could be taken as an indicator of official accountability (Muralidharan et al., 2017), is correlated with lower cheating.

These results should not, of course, be taken as evidence that any of the factors *cause* lower cheating: each of these factors is associated with several possible confounders and with each other. For example, private schools have fewer girls, larger class sizes, differ in the characteristics of their teaching staff and are much more likely to be in urban areas. Rather, these are intended to highlight only that mean differences across groups and even basic associations may look very different across the two sets of assessments, a result that is particularly important for any policy uses of similar data such as creating school league tables or targeting resources.

[31]Schools did not know they would be part of a retest more than 1-2 days in advance at most. The anonymity of results was guaranteed and schools also had the option to refuse consent. However, since this exercise was being conducted in February, which is also the peak pre-exam preparation and revision period in the school year, we cannot rule out that teachers may have used the initial assessment for revision anyway. This is not a concern in the tablet testing arm since no physical question papers were left.

4 Discussion

The they are very unreliable for inferring the level of student achievement although they do contain some useful information about the relative ranks of schools and students. This has important implications for potential uses of such data.

The first thing to note is that it is the *absolute* level of achievement which is the primary policy target in this setting, and in international policy goals. One consequence of the severe cheating is that these data severely understate the 'learning crisis', a phenomenon of low absolute achievement. Put bluntly, they fail to even diagnose the problem they were meant to fix and thus expectations by the World Bank or governments, that these assessments will serve to catalyze action to focus on low achievement, seem optimistic.

That said, the *ordinal* information still retained in these measures could, in principle, still be adequate for many policy purposes. For example, Barlevy and Neal (2012) show how incentive-compatible performance pay mechanisms for teachers can be set up with only ordinal information. Interventions that provide information only on relative achievement or target accountability based on a school or student's position in the overall distribution would also be feasible. Unfortunately, the results also suggest much caution for these uses. For one, since the severity of cheating here does vary by observable characteristics of schools (Table 6), interventions relying on ordinal information alone could also be distorted.

The much larger concern, though, is that even this ordinal information may not survive the addition of high stakes. That cheating is so pervasive, even in the absence of formal incentives, suggests that the costs of cheating, for teachers and students, are very low – and so, while I cannot directly speak to what would happen if the stakes emphasized relative ranks instead, it is plausible that such a switch would distort the ordinal information also. Indeed, one interpretation of the pattern I find in Madhya Pradesh – that levels are severely distorted but ordinal ranks appear not to be — is that this merely reflects the fact that all administrative attention was related to levels and not the ranking of schools. This general principle, that the use of a statistic for policy-making may corrupt existing statistical regularities, has long been acknowledged in both education ('Campbell's law', (Campbell, 1979)) and economic policy ('Goodhart's law', (Goodhart, 1984)). Overall, I interpret the results as suggesting that any use of these test score data as a basis for large interventions will require simultaneous enabling reforms to ensure data integrity, the costs and feasibility of which need to be explicitly accounted for in deriving policy implications when evaluating these interventions.[32]

[32]Thus, for example, meta-analyses of cost-effectiveness of educational interventions, such as presented by Kremer et al. (2013), which purport to inform policy choices, should include not just costs such as the monetary cost of bonuses paid to teachers but also the costs involved in setting up a non-manipulable testing regime.

These results are also sobering for assessments of the potential for using such data for research. Manipulation here poses much more pernicious problems than classical measurement error. Not only is the level of achievement distorted, there is suggestive evidence that such manipulation has differed over time (Table 2), that it varies across students over the achievement distribution (Figure 2), and it varies by observable characteristics of schools (Table 6). It is unlikely that all correlates of manipulation can be adequately captured in large-scale data and it is also plausible that some of these interact with whatever policy or attribute that the researcher is trying to study. Thus, although they may still be useful in specific settings — such as in evaluating pre-treatment differences between schools or students or, subject to some validation, even as supplementary outcome measures — using these assessments as the mainstay of education research is likely to be inadvisable even with low stakes.

The results also present constructive avenues for reform. The first of these, in the spirit of Neal (2013), relates to the importance of multiple measures. Although education systems increasingly prioritize universal census-based assessments, as also seen in National Education Policy(Government of India, 2019), my results suggest that measuring the level and trends in student achievement may be better-served by sample surveys implemented by external organizations with independent proctoring, as with the audit studies here.[33] Such data would also serve as an external benchmark to assess the reliability of administrative data series.

Sample-based data are not, however, adequate for all purposes: many uses need reliable information on *all* units. The generality of cheating I find suggests the need for policies which raise the costs of cheating across-the-board. These appear effective here, both in MP with improved test design and in AP with tablet based assessments, even when implemented entirely by government officials. It is likely that such approaches are more promising, both in effectiveness and scalability, than a primary focus on detecting and punishing individual cheaters, which is perhaps more suitable in settings with lower prevalence (see e.g. Jacob and Levitt, 2003).

Tablet-based assessments appear to eliminate distortion and the intervention I evaluate may provide a useful template for future scale-ups, although there are important further considerations. The first relates to the marginal costs incurred in tablet-based testing. In the A.P. experiment, approximately 3500 tablets were used to test Grade 4 students in the whole district over 10 working days. Under reasonable assumptions, it may be possible to reduce the tablet-related costs to $\sim$0.5 dollars per tested student.[34] Importantly, these are *not* incremental

[33]In contrast to Neal (2013), who stresses the difference in incentives between assessments, this recommendation follows due to constraints to monitoring capacity: it is possible to carry out independent tests in random samples of schools without manipulation but external monitoring in all schools is infeasible.

[34]The marginal cost of testing additional students in the same school on the day of the school visit is low. Extending this testing period, with staggered testing across schools, it would be possible to test multiple grades without needing to significantly increase the number of tablets. I assume that 5000 tablets would suffice for testing around 200,000 students over a 10-week period. Testing is possible with low-specification tablets, which

costs but rather displace substantial costs currently incurred in paper-based tests, especially in effort but also money, related to grading of question papers by teachers, data entry and record-keeping of paper-based tests, and printing costs. Such tests may also make assessments more informative by allowing adaptive testing[35] and, with more granular data, make it easier to detect suspicious responses. Thus, even within current education budgets in LMICs, the adoption of tablet-based tests at a large scale may be feasible and cost-effective.[36]

The second, and important, caveat relates to long-run effects: as a one-off trial, the experiment here does not reflect any long-term adaptation by agents. Such responses have been shown to undermine previous reforms. For instance, Banerjee et al. (2008) and Dhaliwal and Hanna (2017) document how technology-aided public sector interventions to reduce health worker absenteeism were undermined shortly after introduction, with the implicit agreement of higher levels of bureaucracy. Such complete reversal is not universal, though: Muralidharan et al. (2016) present an example of a reform that persisted and, in a different state, Banerjee et al. (2016) document a temporary reversal of the corruption-reducing intervention they study followed by a nationwide scale-up. The long-term sustainability and efficacy of tablet-based tests remains an open question.[37] An important message is that, following *any* such reforms, data integrity will need to be evaluated repeatedly in representative samples.

5 Conclusions

This paper has documented two sets of results. First, that cheating in current census-based assessments in schools severely overstates average achievement. The resulting data are unreliable bases for policy and are unlikely to serve effectively as pre-requisites for remedying the 'learning crisis' in LMICs. Rather, in their existing form, they merely provide a stark illustration of 'isomorphic mimicry' at work where weak organizational systems adopt the *form* of succesful practices, but fail to replicate their *function* (DiMaggio and Powell, 1983; Pritchett et al., 2013). Second, however, substantial reductions in distortion may be possible. Multiple test booklets and external grading in M.P., and tablet-based assessments in A.P. reduce cheating substantially even though implemented at scale by government officials. Given the willingness of governments

currently cost about USD 60. Assuming these have a three-year depreciation period, this translates into a per-test cost of about 50 cents.

[35]This is particularly important since students are often very far behind grade-appropriate levels and thus tests tied closely to official syllabi alone may miss even substantial learning gains (Muralidharan et al., 2019).

[36]This points to a different use of technology in education systems, much closer to the literature on public sector corruption, than the more widely-studied case of using technology to relax constraints to effective pedagogy (e.g. Barrow, Markman and Rouse 2009 and Muralidharan, Singh and Ganimian 2019).

[37]This concern is not, of course, specific to this intervention or even just to large-scale policy experiments. See, for instance, Jayaraman, Ray and De Véricourt (2016) for an example of how adaptation by agents reverses short-term conclusions about the productivity effects of payment schedules for workers.

to invest in such census-based assessments, such interventions may have high rewards both for policy and for research.

While measuring student achievement is an important application in itself, the challenge presented by distortion in administrative data is much broader. In many sectors such as health, public finance, or social security, administrative data are likely to be as susceptible to manipulation and are as central for planning, targeting and policy implementation.[38] As in the specific case analysed in this paper, the corruptibility of information severely limits the uses to which administrative data could be put and directly constrains the efficient provision of basic services. In this respect, administrative data, both in its scope and reliability, forms a key part of the basic infrastructure of state capacity for implementation.[39]

Although the precise way to improve such data is likely to differ across settings, it is possible that the core principles of designing more robust administrative data systems presented in this paper – sample-based audits of administrative data to provide an external benchmark, and the use of technology to limit the possibility of corruption – may also generalize across sectors. As such, like interventions such as performance-related pay or the improved use of communication technology for service delivery, such interventions may be rewarding to experiment with in different settings. Attempts in this direction, as with all large public interventions, should be informed and disciplined by rigorous evaluations in representative populations at scale.

The overarching message of this paper, therefore, is to suggest caution for strategies that emphasize "data-driven" approaches to policy reform without considering the provenance of the data themselves. The creation of robust data systems with standardized measurements and individual-level data, has historically been a major (and often contentious) challenge in developing administrative capacity across the world. As this paper highlights, this challenge remains relevant in developing countries today, not only in areas where incentives are very clear-cut, such as taxation and the standardized measurement of property, but potentially even in settings where the measurement is not directly tied to formal incentives. Remedying such distortion at scale remains a substantial task for administrative systems.

[38]For instance, similar concerns have been raised in India about government data on the provision of basic sanitation, an area of high priority in public health and one of the signature campaigns of the present national administration. Administrative data declaring villages as "open-defecation free" are reported to be substantially overstated, despite in principle having a detailed process for verification by multiple administrative authorities (Agarwal, 2019).

[39]Of course, improving the reliability of administrative data may not be desirable in all settings. The strong presumption in this paper is that of a setting where the the interests of policymakers and the public are aligned with each other and increases in state capacity are welfare-improving. This is not, however, a universal case (see Scott 1998 for several such examples). This deep ambiguity attaches to *all* interventions which aim to improve state capacity, and especially so in settings with weaker institutions.

References

Afridi, Farzana, Bidisha Barooah, and Rohini Somanathan, "Improving learning outcomes through information provision: Experimental evidence from Indian villages," *Journal of Development Economics*, 2018.

Agarwal, Kabir, "Government Data Proves We Shouldn't Believe India Is 'Open Defecation Free," *The Wire, India*, 2019.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja, "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets," *American Economic Review*, 2017, *107* (6), 1535–63.

Angrist, Joshua D, Erich Battistin, and Daniela Vuri, "In a small moment: Class size and moral hazard in the Italian Mezzogiorno," *American Economic Journal: Applied Economics*, 2017, *9* (4), 216–49.

Banerjee, Abhijit, Esther Duflo, Clement Imbert, Santhosh Mathew, and Rohini Pande, "E-governance, accountability, and leakage in public programs: Experimental evidence from a financial management reform in india," Technical Report, National Bureau of Economic Research 2016.

Banerjee, Abhijit V, Esther Duflo, and Rachel Glennerster, "Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system," *Journal of the European Economic Association*, 2008, *6* (2-3), 487–500.

Barlevy, Gadi and Derek Neal, "Pay for percentile," *American Economic Review*, 2012, *102* (5), 1805–31.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse, "Technology's edge: The educational benefits of computer-aided instruction," *American Economic Journal: Economic Policy*, 2009, *1* (1), 52–74.

Battistin, Erich, Michele De Nadai, and Daniela Vuri, "Counting rotten apples: Student achievement and score manipulation in Italian elementary Schools," *Journal of Econometrics*, 2017, *200* (2), 344–362.

Bertoni, Marco, Giorgio Brunello, and Lorenzo Rocco, "When the cat is near, the mice won't play: The effect of external examiners in Italian schools," *Journal of Public Economics*, 2013, *104*, 65–77.

Bharadwaj, Prashant, Katrine Vellesen Løken, and Christopher Neilson, "Early life health interventions and academic achievement," *American Economic Review*, 2013, *103* (5), 1862–91.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Justin Sandefur et al., "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 2018, *168*, 1–20.

Borcan, Oana, Mikael Lindahl, and Andreea Mitrut, "Fighting corruption in education: What works and who benefits?," *American Economic Journal: Economic Policy*, 2017, *9* (1), 180–209.

Campbell, Donald T, "Assessing the impact of planned social change," *Evaluation and program planning*, 1979, *2* (1), 67–90.

Cattaneo, Matias D, Richard K Crump, Max H Farrell, and Yingjie Feng, "On binscatter," *arXiv preprint arXiv:1902.09608*, 2019.

Chay, Kenneth Y, Patrick J McEwan, and Miguel Urquiola, "The central role of noise in evaluating interventions that use test scores to rank schools," *American Economic Review*, 2005, *95* (4), 1237–1258.

Dee, Thomas, Will Dobbie, Brian A Jacob, and Jonah E Rockoff, "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations," *American Economic Journal: Applied Economics*, 2019, *11* (3), 382–423.

Dhaliwal, Iqbal and Rema Hanna, "The devil is in the details: The successes and limitations of bureaucratic reform in India," *Journal of Development Economics*, 2017, *124*, 1–21.

Diamond, Rebecca and Petra Persson, "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests," *NBER Working Paper*, 2016, (w22207).

DiMaggio, Paul J and Walter W Powell, "The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields," *American Sociological Review*, 1983, pp. 147–160.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, "Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India," *The Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

Figlio, David, Krzysztof Karbownik, and Kjell G Salvanes, "Education research and administrative data," in "Handbook of the Economics of Education," Vol. 5, Elsevier, 2016, pp. 75–138.

Fujiwara, Thomas, "Voting technology, political responsiveness, and infant health: Evidence from Brazil," *Econometrica*, 2015, *83* (2), 423–464.

Gilligan, Daniel O, Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M Lucas, and Derek Neal, "Educator incentives and educational triage in rural primary schools," Technical Report, National Bureau of Economic Research 2018.

Glewwe, P and K Muralidharan, "Improving Education Outcomes in Developing Countries – Evidence, Knowledge Gaps, and Policy Implications," *Handbook of the Economics of Education*, 2016, *5.*

Goodhart, Charles AE, "Problems of monetary management: the UK experience," in "Monetary Theory and Practice," Springer, 1984, pp. 91–121.

Government of India, *Draft National Education Policy 2019*, Ministry of Human Resource Development, Government of India, New Delhi, 2019.

Hinnerich, Björn Tyrefors and Jonas Vlachos, "The impact of upper-secondary voucher school attendance on student achievement. Swedish evidence using external and internal evaluations," *Labour Economics*, 2017, *47*, 1–14.

Hsieh, Chang-Tai and Miguel Urquiola, "The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program," *Journal of public Economics*, 2006, *90* (8-9), 1477–1503.

Imbert, Clement and John Papp, "Labor market effects of social programs: Evidence from india's employment guarantee," *American Economic Journal: Applied Economics*, 2015, *7* (2), 233–63.

Jacob, Brian A and Steven D Levitt, "Rotten apples: An investigation of the prevalence and predictors of teacher cheating," *The Quarterly Journal of Economics*, 2003, *118* (3), 843–877.

Jacob, Brian and Jesse Rothstein, "The measurement of student ability in modern assessment systems," *Journal of Economic Perspectives*, 2016, *30* (3), 85–108.

Jayaraman, Rajshri, Debraj Ray, and Francis De Véricourt, "Anatomy of a contract change," *American Economic Review*, 2016, *106* (2), 316–58.

Kremer, Michael, Conner Brannen, and Rachel Glennerster, "The challenge of education and learning in the developing world," *Science*, 2013, *340* (6130), 297–300.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin, "Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools," *mimeo.*, 2020.

Lewis-Faupel, Sean, Yusuf Neggers, Benjamin A Olken, and Rohini Pande, "Can electronic procurement improve infrastructure provision? Evidence from public works in India and Indonesia," *American Economic Journal: Economic Policy*, 2016, *8* (3), 258–83.

Linden, Leigh L and Gauri Kartini Shastry, "Grain inflation: Identifying agent discretion in response to a conditional school nutrition program," *Journal of Development Economics*, 2012, *99* (1), 128–138.

Loyalka, Prashant, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi, "Pay by design: Teacher performance pay design and the distribution of student achievement," *Journal of Labor Economics*, 2019, *37* (3), 621–662.

Martinelli, César and Susan Wendy Parker, "Deception and misreporting in a social program," *Journal of the European Economic Association*, 2009, *7* (4), 886–908.

_ , Susan W Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo, "Cheating and incentives: Learning from a policy experiment," *American Economic Journal: Economic Policy*, 2018, *10* (1), 298–325.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani, "Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania," *The Quarterly Journal of Economics*, 2019, *134* (3), 1627–1673.

Mizala, Alejandra and Miguel Urquiola, "School markets: The impact of information approximating schools' effectiveness," *Journal of Development Economics*, 2013, *103*, 313–335.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro Ganimian, "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *The American Economic Review*, 2019, *Forthcoming.*

_ **and** _ , "Improving Public Sector Management at Scale: Experimental Evidence on School Governance in India," *mimeo.*, 2019.

_ **and Paul Niehaus**, "Experimentation at scale," *Journal of Economic Perspectives*, 2017, *31* (4), 103–24.

_ **and Venkatesh Sundararaman**, "The impact of diagnostic feedback to teachers on student learning: experimental evidence from India," *The Economic Journal*, 2010, *120* (546), F187–F203.

_ **and** _ , "Teacher performance pay: Experimental evidence from India," *Journal of Political Economy*, 2011, *119* (1), 39–77.

_ **and** _ , "The aggregate effect of school choice: Evidence from a two-stage experiment in India," *The Quarterly Journal of Economics*, 2015, *130* (3), 1011–1066.

_ **, Jishnu Das, Alaka Holla, and Aakash Mohpal**, "The fiscal cost of weak governance: Evidence from teacher absence in India," *Journal of Public Economics*, 2017, *145*, 116–135.

_ **, Paul Niehaus, and Sandip Sukhtankar**, "Building state capacity: Evidence from biometric smartcards in India," *American Economic Review*, 2016, *106* (10), 2895–2929.

_ **,** _ **,** _ **, and Jeffrey Weaver**, "Improving Last-Mile Service Delivery using Phone-Based Monitoring," Technical Report, National Bureau of Economic Research, Inc 2018.

Neal, Derek, "The consequences of using one assessment system to pursue two objectives," *The Journal of Economic Education*, 2013, *44* (4), 339–352.

Neilson, Christopher, Claudia Allende, and Francisco Gallego, "Approximating the Equilibrium Effects of Informed School Choice," 2019.

Niehaus, Paul and Sandip Sukhtankar, "The marginal rate of corruption in public programs: Evidence from India," *Journal of Public Economics*, 2013, *104*, 52–64.

NITI Aayog, *Social Sector Service Delivery: Good Practice Resource Book*, NITI Aayog, Government of India, New Delhi, 2016.

Olken, Benjamin A, "Monitoring corruption: evidence from a field experiment in Indonesia," *Journal of Political Economy*, 2007, *115* (2), 200–249.

Pratham, *Annual Status of Education Report 2016*, Pratham, New Delhi, 2017.

_ , *Annual Status of Education Report 2018*, Pratham, New Delhi, 2019.

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

Reinikka, Ritva and Jakob Svensson, "Local capture: evidence from a central government transfer program in Uganda," *The Quarterly Journal of Economics*, 2004, *119* (2), 679–705.

Scott, James C, *Seeing like a state: How certain schemes to improve the human condition have failed*, Yale University Press, 1998.

Singh, Abhijeet, "Private school effects in urban and rural India: Panel estimates at primary and secondary school ages," *Journal of Development Economics*, 2015, *113*, 16–32.

_ , "Learning More With Every Year: School Year Productivity and International Learning Divergence," *Journal of the European Economic Association*, 2019, *Forthcoming.*

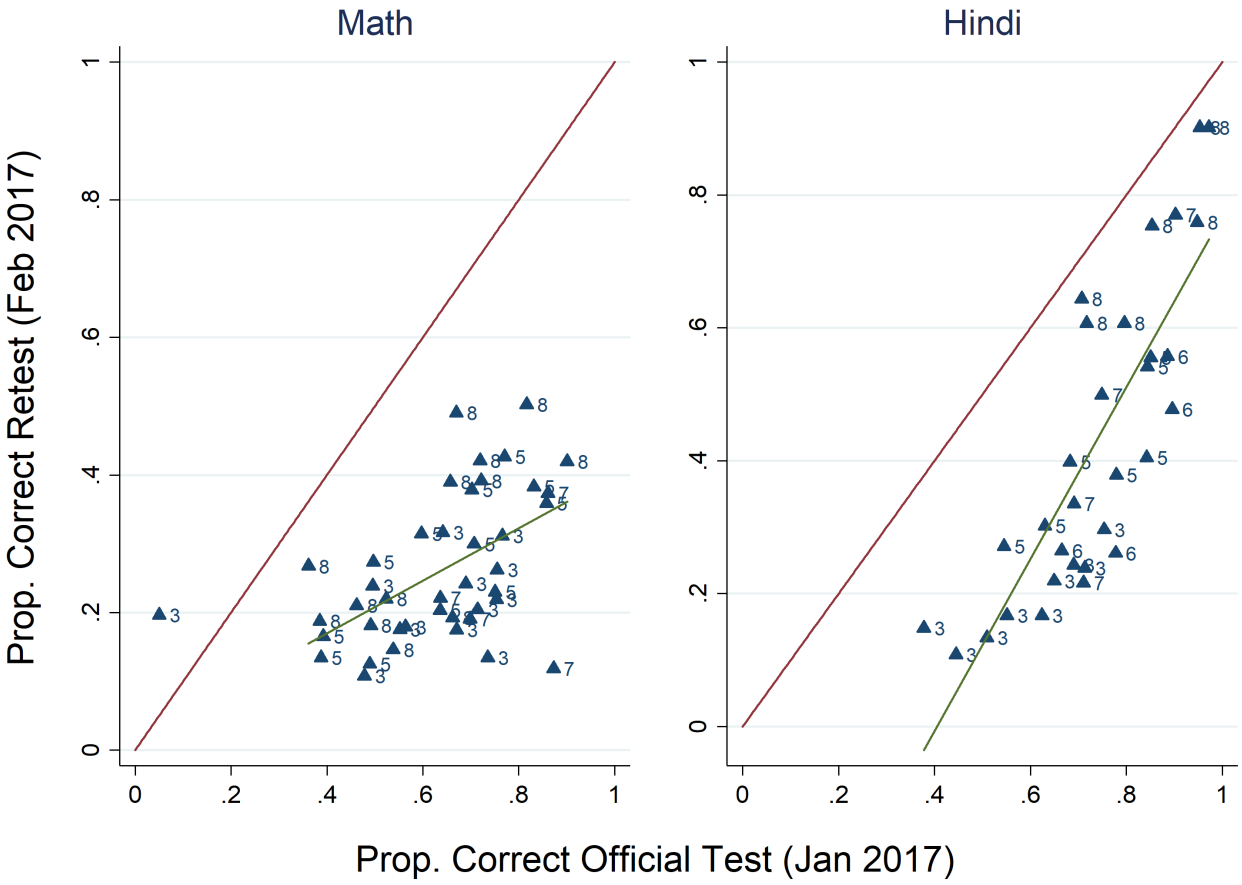
Vivalt, Eva, "How much can we generalize from impact evaluations?," *Journal of the European Economic Association*, forthcoming.

Wikström, Christina and Magnus Wikström, "Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools," *Economics of Education Review*, 2005, *24* (3), 309–322.

World Bank, *World Development Report 2018: Learning to realize education's promise*, The World Bank, Washington DC, 2018.

Figures

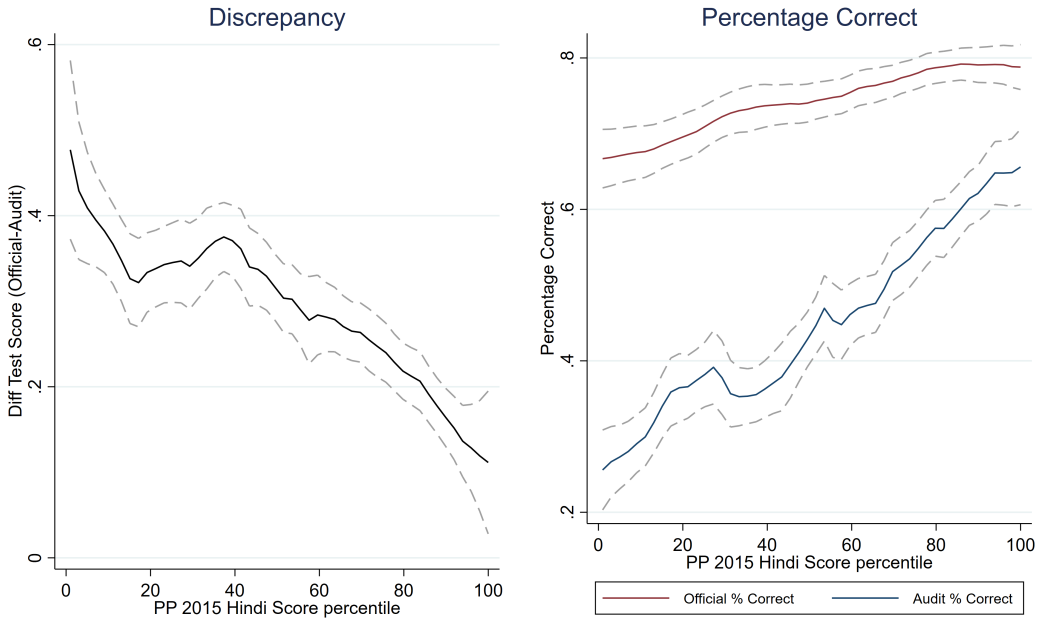
Figure 1: Comparing item-level data from official tests and retest audit



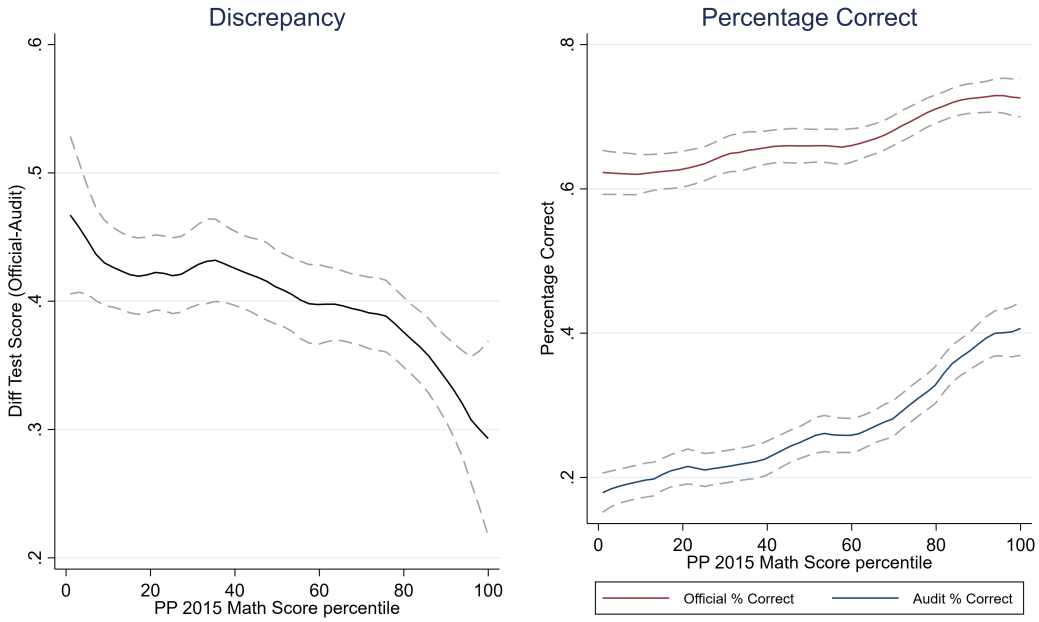
Note: Each dot in this figure is an individual multiple-choice test question and compares the proportion of students who are reported to have correctly answered in the Pratibha Parv assessment (Jan 2017) with the percentage correctly answered in the audit (Feb 2017). The marker label indicates the grade in which the question was administered.

Figure 2: Discrepancy over the achievement distribution

(a) Hindi

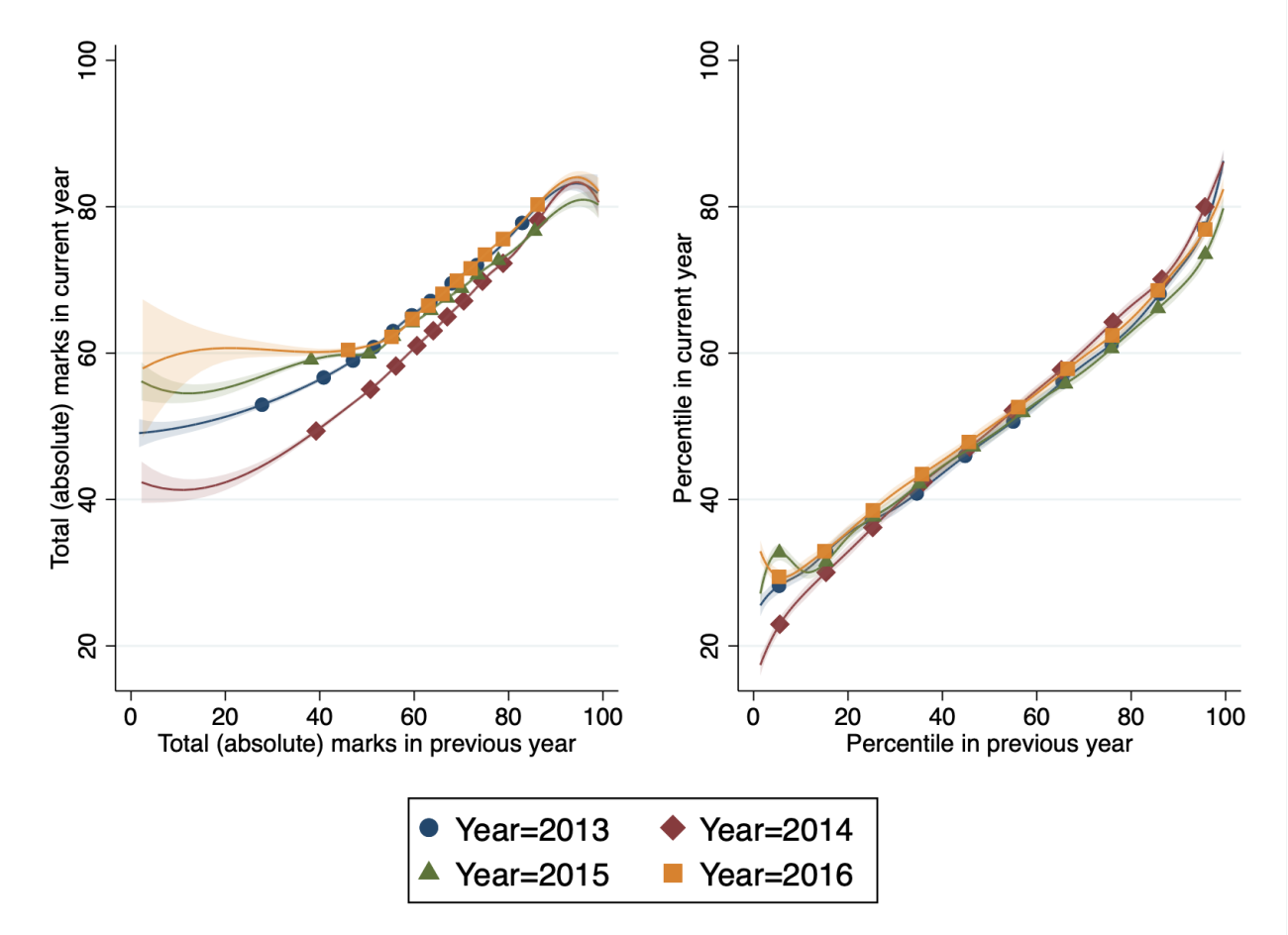


(b) Math



Note: Discrepancy is defined as the difference between proportion correctly answered in the official assessment and the retest for test items which are common across both assessments. Percentiles are defined over the test score in the previous academic year. The left panel shows the variation of mismatch over the percentiles of achievement in the previous year's test. The right panel shows the conditional mean of the percentage correct, on the same common items, in the official test (PP) and the audit (EL) across the same percentiles. The distance between the two curves provides the mismatch measure.

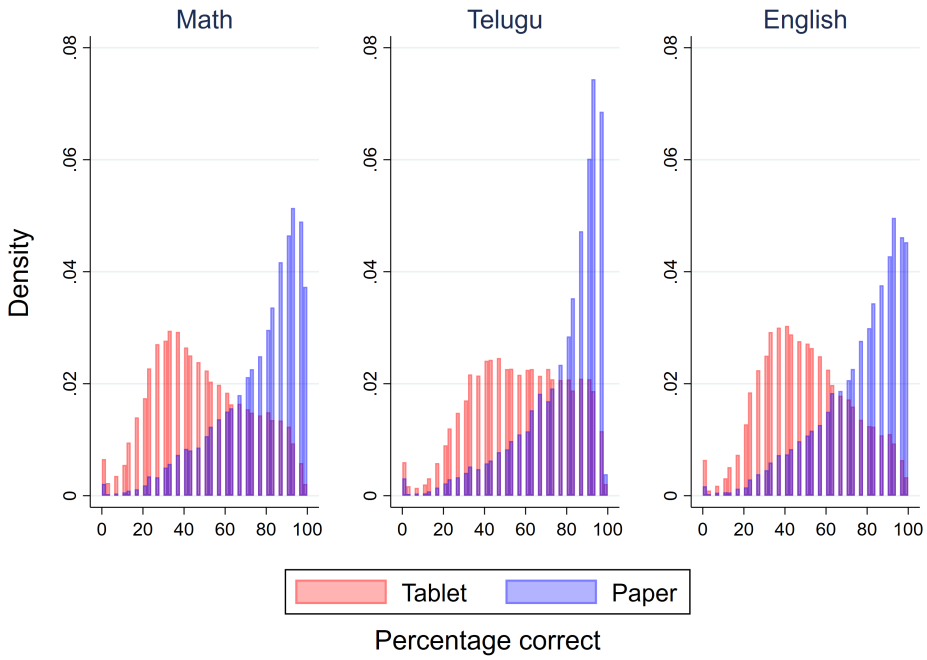
Figure 3: Stability of ordinal ranks at the school level



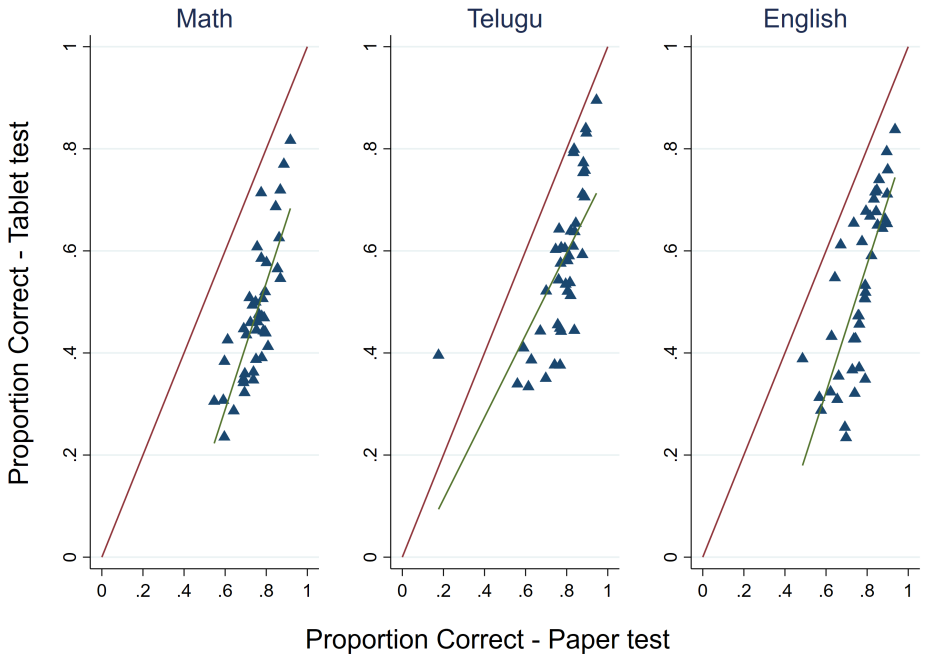
Note: This figure relates absolute marks at the school level (left panel) in the official data, and the percentile in the within-year distribution (right panel), in one year to the previous year, using the procedure of Cattaneo et al. (2019). Markers denote average in each decile, the line shows a third-order polynomial in each bin, and the confidence band shows 95% confidence level. The relationship in absolute marks has changed substantially over years, especially at the bottom end. However, the relationship between percentile ranks has been stable for the entire 5-year period.

Figure 4: Test scores in paper and tablet tests

(a) Test score distribution in tablet and paper tests – Student level

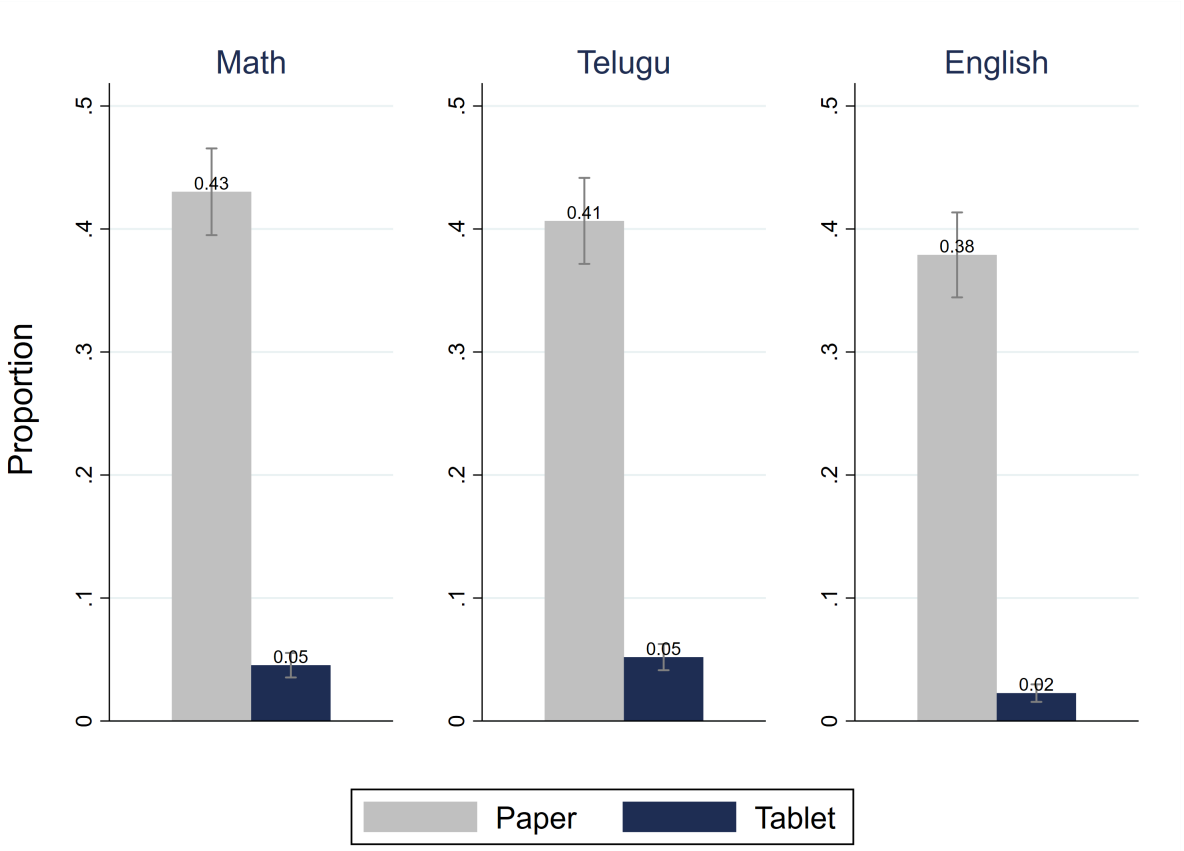


(b) Proportion correct in tablet and paper tests – Item level



Note: Panel (a) shows the distribution of student-wise aggregate test scores (percentage correct) for students who were tested on the same question papers using tablet-based or paper-based assessments. Panel (b) shows the percentage of students correctly answering the same question (denoted by triangle) across the paper and tablet tests. Schools were randomly assigned to be tested on paper or tablet assessments.

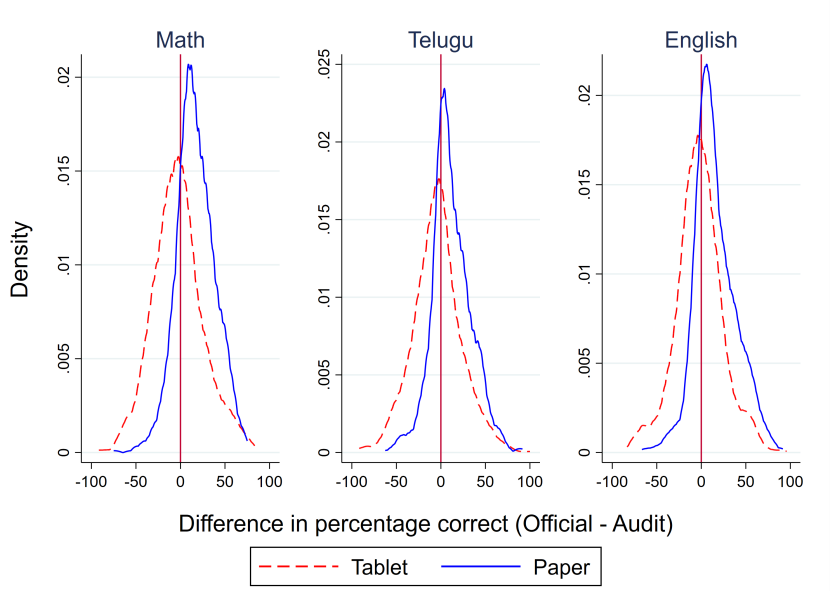
Figure 5: Proportion of schools flagged as potentially cheating



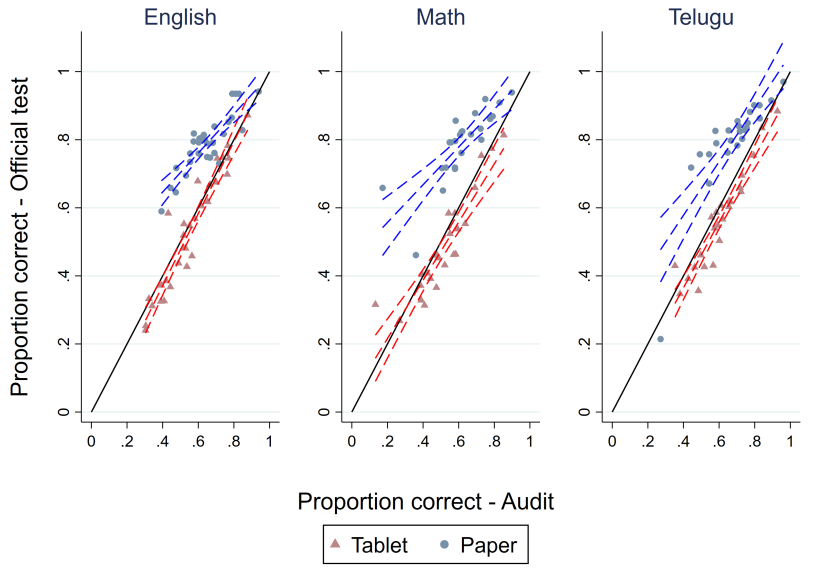
Note: This figure shows the proportion of schools in the paper and tablet testing arms which are flagged as having potentially cheated based on the procedure in Angrist et al. (2017). This procedure identifies, at the classroom level, anomalous response patterns using item level data; please see Appendix C for details. Whereas between 38-43% of classrooms with paper-based testing are flagged, this figure is only around 2-5% in tablet based tests.

Figure 6: Correspondence of the retest with the official paper and tablet tests

(a) Difference in total percentage correct (Official - Audit)



(b) Item-level percentage correct in official test and audit



Note: Panel (a) shows the difference in percentage correct in the official test and the retest audit on common items at the student level. Panel (b) shows the difference in the proportion of students answering the same question correctly in the the official test (vertical axis) and the retest (horizontal axis). Whereas average deviation in responses between the tablet tests and the retest is very small, student achievement is overstated in paper-based assessments that were proctored internally by school teachers.

Tables

Table 1: Sample characteristics of surveyed schools in Madhya Pradesh

	Whole state mean/sd	Study districts mean/sd	Sample mean/sd
School-level characteristics			
Enrolment in Elementary School	65.77	60.41	52.95
	(57.25)	(50.82)	(35.00)
No. of teachers	2.62	2.48	2.33
	(1.68)	(1.50)	(1.23)
Proportion of female teachers	0.28	0.28	0.30
	(0.33)	(0.34)	(0.35)
Pupil-teacher ratio	28.19	27.69	25.66
	(23.17)	(21.68)	(19.92)
Rural	0.94	0.95	0.94
	(0.24)	(0.23)	(0.23)
Observations	114286	24183	283
School-level test scores			
Pratibha Parv school score	68.98	68.82	68.55
	(11.63)	(10.60)	(11.18)
A Grade School	0.31	0.28	0.29
	(0.46)	(0.45)	(0.45)
B Grade School	0.49	0.52	0.48
	(0.50)	(0.50)	(0.50)
C Grade School	0.18	0.18	0.20
	(0.38)	(0.39)	(0.40)
D Grade School	0.02	0.01	0.02
	(0.15)	(0.12)	(0.16)
E Grade School	0.01	0.00	0.00
	(0.07)	(0.04)	(0.00)
Observations	111138	23741	283

The table presents mean values of observable school-level characteristics using year 2016 DISE Data and of school-level test scores using 2016-17 official Pratibha Parv test score Data for schools in Madhya Pradesh. The first column shows data for all government-run schools with grades 1-8. The second column shows data for the study population which includes all government-run schools with grades 1-8 from five districts in the Bhopal region (Bhopal, Raisen, Rajgarh, Sehore and Vidisha) and tribal blocks from five districts in the Indore region (Alirajpur, Barwani, Dhar, Jhabua, and Khargone). The third column contains the sample of schools for whic item-level data was matched for the indepndent and official data. Standard deviations are reported in parentheses.

Table 2: Evolution of aggregate scores in Pratibha Parv

	2012	2013	2014	2015	2016	Average
Total Marks (Mean)	54.51	63.15	63.14	66.52	68.98	63.22
(Standard Deviation)	(18.563)	(15.680)	(14.320)	(11.358)	(11.625)	(15.381)
% of schools with grade A	0.10	0.22	0.21	0.22	0.31	0.21
% of schools with grade B	0.30	0.43	0.43	0.53	0.49	0.43
% of schools with grade C	0.35	0.25	0.27	0.22	0.18	0.26
% of schools with grade D	0.14	0.06	0.06	0.03	0.02	0.06
% of schools with grade E	0.11	0.03	0.03	0.01	0.01	0.04
% of students with grade A	0.18	0.25	0.25	0.23	0.23	0.23
% of students with grade B	0.27	0.32	0.31	0.33	0.33	0.31
% of students with grade C	0.30	0.26	0.26	0.34	0.35	0.30
% of students with grade D	0.10	0.06	0.06	0.04	0.07	0.06
% of students with grade E	0.14	0.12	0.12	0.03	0.01	0.09
Total Enrollment	67.29	81.64	75.95	63.20	63.09	70.29
Number of Schools	114005	114308	113598	112346	111138	113091

Source: Pratibha Parv administrative data. Values in parentheses correspond to the standard deviation of Pratibha Parv marks at the school level.

Table 3: Reduction of mismatch with improved exam procedures

VARIABLES	(1)	(2)	(3)	(4)
	Diff in proportion correctly answered (PP-Audit)			
	Math	Math	Hindi	Hindi
Treatment	-0.235***	-0.235***	-0.264***	-0.274***
	(0.0468)	(0.0469)	(0.0254)	(0.0247)
Constant	0.541***	0.541***	0.358***	0.360***
	(0.0534)	(0.0210)	(0.0288)	(0.00653)
School FE		Yes		Yes
Observations	1,081	1,080	1,526	1,525
R-squared	0.108	0.378	0.095	0.249
Average mismatch	0.436	0.436	0.288	0.288
No. schools	49	48	49	48

Note: *** p<0.01, ** p<0.05, * p<0.1 Robust standard errors, clustered at school level, in parentheses. The dependent variable is defined as the difference between the proportion correctly answered in the Pratibha Parv assessment score and the retest audit, on the questions which were common between both assessments. Treatment is an indicator variable that takes the value of 1 for Grade 8, which had multiple sets and external grading, and is 0 for Grades 6 and 7. In mathematics, there were no multiple-choice questions that were common between the test and the retest and therefore the estimation sample only includes Grades 7 and 8. Standard errors are clustered at the school level.

Table 4: Balance of observables across paper and tablet testing arms

	Initial Assignment				Final Assignment				Retest Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Variable	Tablet	Paper	Diff	Diff SE	Tablet	Paper	Diff	Diff SE	Tablet	Paper	Diff	Diff SE
Enrolment												
Class 4	18.08	17.11	0.97	(1.34)	17.94	17.40	0.54	(1.31)	20.22	18.22	2.00	(3.71)
Primary (Class I-V)	88.51	84.82	3.70	(6.41)	88.03	85.83	2.20	(6.32)	102.17	96.18	5.98	(19.72)
School Characteristics												
Government or aided	0.80	0.80	-0.01	(0.05)	0.80	0.79	0.01	(0.05)	0.58	0.58	0.00	(0.13)
Private Unaided	0.20	0.20	0.01	(0.05)	0.20	0.21	-0.01	(0.05)	0.42	0.42	-0.00	(0.13)
Rural	0.86	0.85	0.02	(0.07)	0.86	0.85	0.01	(0.07)	0.77	0.87	-0.10	(0.13)
English medium	0.17	0.18	-0.02	(0.05)	0.16	0.18	-0.02	(0.05)	0.27	0.40	-0.13	(0.13)
Telugu medium	0.83	0.82	0.02	(0.05)	0.84	0.82	0.02	(0.05)	0.73	0.60	0.13	(0.13)
Infrastructure												
Num of classrooms	3.60	3.49	0.11	(0.16)	3.62	3.45	0.17	(0.16)	3.67	3.87	-0.20	(0.49)
Num of toilets	2.93	2.73	0.20	(0.18)	2.90	2.81	0.09	(0.18)	3.33	2.83	0.50	(0.53)
Electricity	0.98	0.98	0.00	(0.01)	0.98	0.98	0.01	(0.01)	0.97	0.98	-0.02	(0.03)
Headmaster room	0.24	0.24	-0.00	(0.04)	0.23	0.25	-0.02	(0.04)	0.38	0.35	0.03	(0.11)
Playground	0.57	0.57	-0.00	(0.03)	0.56	0.59	-0.02	(0.03)	0.62	0.70	-0.08	(0.09)
No boundary wall	0.45	0.47	-0.02	(0.03)	0.46	0.46	-0.00	(0.04)	0.42	0.47	-0.05	(0.10)
Inspections												
Visits by BRC	1.92	1.91	0.02	(0.20)	1.93	1.89	0.04	(0.20)	1.92	1.52	0.40	(0.53)
Visits by CRC	3.16	3.06	0.11	(0.37)	3.20	2.98	0.22	(0.36)	2.98	2.27	0.72	(0.85)
Observations	1,685	777	2,462		1,694	768	2,462		60	60	120	

The value displayed for t-tests are the differences in the means across the groups. Standard errors are clustered at variable cluster. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 5: Correspondence of the retest with the official paper and tablet tests

	(1)	(2)	(3)
	Diff. between main test and audit		
VARIABLES	English	Telugu	Math
Paper test	0.169***	0.160***	0.197***
	(0.0300)	(0.0366)	(0.0441)
Constant	-0.0228	-0.0415	-0.0267
	(0.0243)	(0.0345)	(0.0429)
Observations	43,482	39,360	39,600
R-squared	0.021	0.019	0.026

Each observation is at the student level, matched across the core test and the retest. The dependent variable is the difference between having answered correctly in the main test and in the retest (main test - retest). Standard errors are clustered at the cluster level.

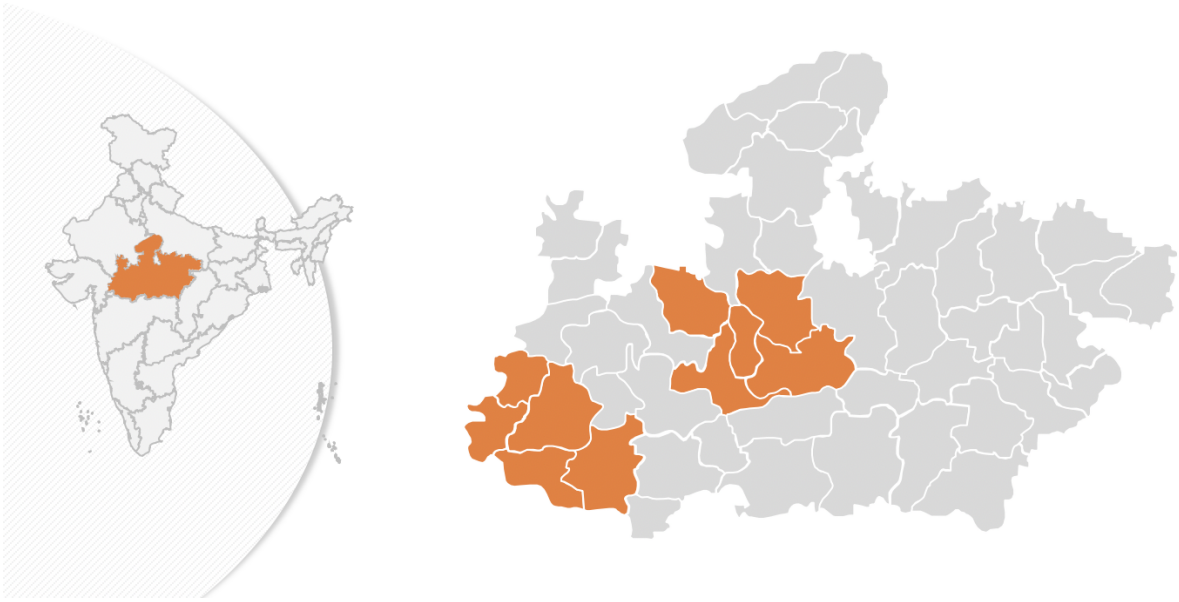
Table 6: Correlates of cheating

	(1) Private	(2) G4 Enrolment	(3) Prop Girls	(4) # Inspections
		Mathematics		
Tablet	-23.94***	-24.70***	-32.57***	-27.48***
	(1.021)	(1.109)	(2.430)	(1.259)
Covariate	3.109**	0.00202	-5.403	0.0587
	(1.386)	(0.0172)	(3.886)	(0.109)
Interaction term	-8.727***	-0.0820***	11.29**	0.0700
	(1.701)	(0.0225)	(4.457)	(0.137)
Constant	74.98***	76.03***	78.53***	75.69***
	(0.871)	(0.932)	(2.049)	(0.940)
Observations	37,660	37,660	37,552	37,660
R-squared	0.314	0.311	0.307	0.307
Student Covariate Mean	0.448	33.73	0.474	4.328
School Covariate Mean	0.250	17.58	0.494	5.030
		Telugu (language)		
Tablet	-17.39***	-17.86***	-26.06***	-19.96***
	(0.901)	(1.053)	(2.231)	(1.143)
Covariate	2.688**	-0.00125	-7.716**	0.134
	(1.261)	(0.0186)	(3.417)	(0.100)
Interaction term	-7.834***	-0.0811***	11.99***	-0.0551
	(1.591)	(0.0231)	(4.057)	(0.124)
Constant	76.92***	77.95***	81.48***	77.12***
	(0.738)	(0.899)	(1.834)	(0.861)
Observations	37,481	37,481	37,370	37,481
R-squared	0.226	0.225	0.219	0.219
Student Covariate Mean	0.446	33.61	0.475	4.334
School Covariate Mean	0.250	17.47	0.493	5.018
		English		
Tablet	-24.31***	-24.38***	-25.75***	-23.60***
	(1.051)	(1.079)	(2.523)	(1.226)
Covariate	6.214***	0.0527**	-3.777	-0.0342
	(1.596)	(0.0225)	(3.952)	(0.0921)
Interaction term	-3.209*	-0.0192	1.651	-0.249**
	(1.888)	(0.0245)	(4.441)	(0.113)
Constant	73.57***	74.10***	77.61***	75.81***
	(0.873)	(0.951)	(2.197)	(0.941)
Observations	37,849	37,849	37,741	37,849
R-squared	0.311	0.308	0.305	0.308
Student Covariate Mean	0.448	33.82	0.474	4.342
School Covariate Mean	0.249	17.61	0.494	5.042

Note: All regressions include block fixed effects. Robust standard errors in parentheses clustered at the cluster level. The dependent variable in all regressions is the subject percentage correct. The regressions are run using a student-level dataset. Student covariate mean refer to the means of the column covariates calculated at the student level. School covariate mean refer to the means of the column covariates calculated at the school level.

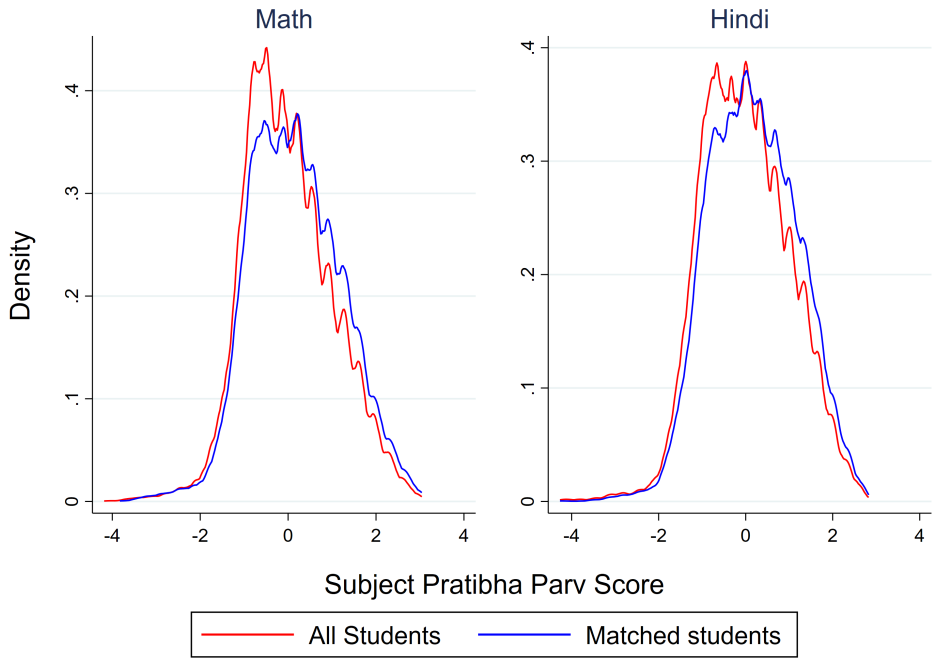
A Additional Figures and Tables

Figure A.1: Sample districts in Madhya Pradesh



Note: The districts highlighted in orange show the setting for field data collection in MP. These comprise the five districts in the Bhopal region and the five districts in the Indore region.

Figure A.2: Comparing matched vs. unmatched students in MP

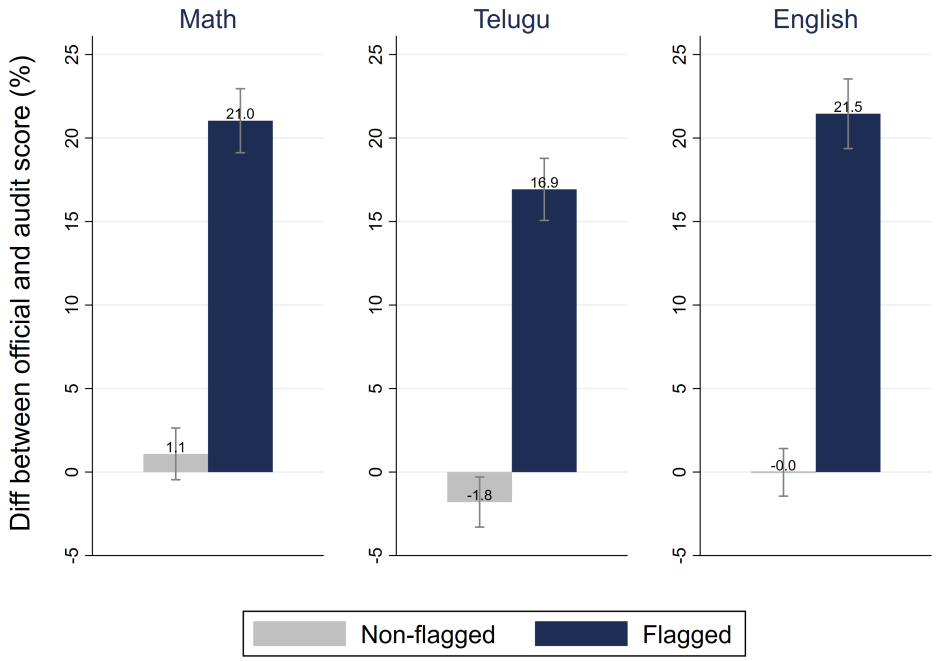


Note: This graph compares the distribution of scores on the official test for students for whom we also have independent assessment data (matched students) to the distribution for the full population of students. Matched students have somewhat higher test scores but the distribution of achievement has substantial overlap and is very similar across the samples. Since being in the matched sample is contingent on being present on the day of the independent test, and well-performing students typically have higher attendance, this moderate difference in achievement of about 0.17 s.d. higher in both subjects is not unexpected.

Figure A.3: Sample district in Andhra Pradesh

Note: This map highlights the setting for the experiment in Andhra pradesh, which was carried out in Prakasam district with Grade 4 students.

Figure A.4: Audit correspondence with official test for flagged and non-flagged schools



Note: This figure compares the difference between the official test score and the audit, which serves as our direct measure of cheating, across the schools which are flagged for potential manipulation using the indirect procedure of Angrist et al. (2017). As can be seen, there is considerable agreement across the two metrics. There is very little evidence of any disagreement on average in schools that are not flagged by the indirect analysis, whereas the difference is pronounced (17-21 pp) in the schools flagged as cheating.

Figure A.5: Comparing performance in the retest across the tablet and paper arms

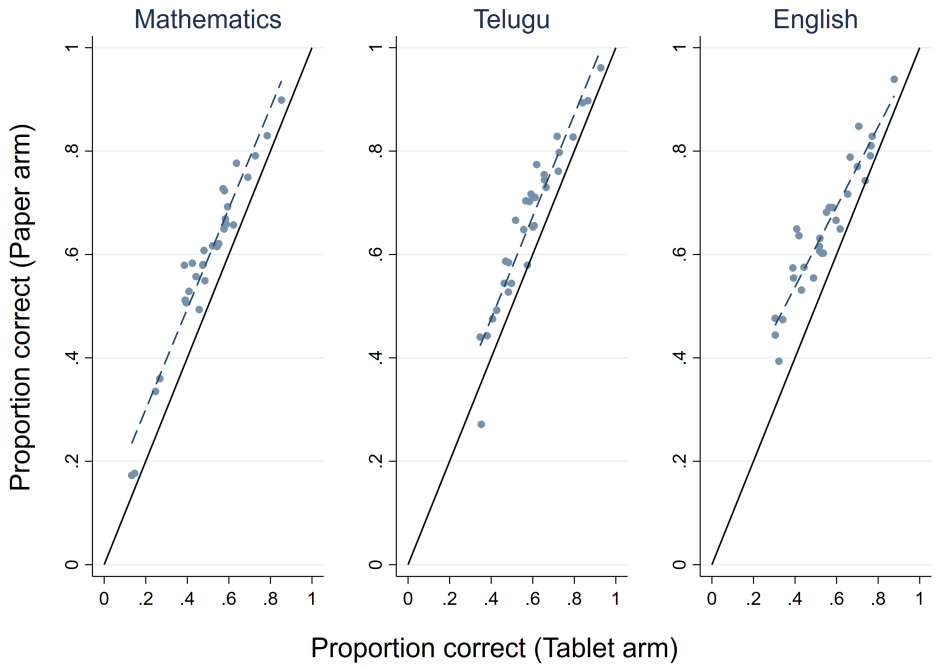


Table A.1: Balance of observables across administered tablet and paper testing arms

	Administered Assignment				Retest Sample			
Variable	Tablet (1)	Paper (2)	Diff (3)	Diff SE (4)	Tablet (5)	Paper (6)	Diff (7)	Diff SE (8)
Enrolment								
Class 4	17.89	17.35	0.54	(1.33)	21.00	18.22	2.78	(3.88)
Primary (Class I-V)	87.85	85.76	2.09	(6.29)	105.79	96.18	9.61	(20.80)
School Characteristics								
Government or aided	0.80	0.79	0.01	(0.04)	0.60	0.58	0.01	(0.13)
Private Unaided	0.20	0.21	-0.01	(0.04)	0.40	0.42	-0.01	(0.13)
Rural	0.86	0.84	0.02	(0.06)	0.75	0.85	-0.10	(0.14)
English medium	0.16	0.19	-0.03	(0.05)	0.25	0.38	-0.14	(0.13)
Telugu medium	0.84	0.81	0.03	(0.05)	0.75	0.62	0.14	(0.13)
Infrastructure								
Num of classrooms	3.62	3.43	0.19	(0.17)	3.74	3.87	-0.13	(0.52)
Num of toilets	2.89	2.83	0.06	(0.17)	3.44	2.83	0.61	(0.56)
Electricity	0.99	0.98	0.01	(0.01)	0.98	0.98	-0.00	(0.02)
Headmaster room	0.23	0.25	-0.02	(0.04)	0.40	0.35	0.05	(0.11)
Playground	0.56	0.59	-0.02	(0.03)	0.61	0.70	-0.09	(0.09)
No boundary wall	0.46	0.46	-0.00	(0.04)	0.40	0.47	-0.06	(0.11)
Inspections								
Visits by BRC	1.93	1.88	0.06	(0.20)	1.96	1.52	0.45	(0.54)
Visits by CRC	3.21	2.97	0.24	(0.35)	3.04	2.27	0.77	(0.85)
Observations	1,683	760	2,443		57	60	117	

The value displayed for t-tests are the differences in the means across the groups. Standard errors are clustered at variable cluster. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

B Direct observations and qualitative interviews about Pratibha Parv

B.1 Direct observations of testing

In both the 2016-17 and 2017-18 school years, we conducted direct observations of the administration of the Pratibha Parv assessments. The purpose of the observations was to inform our later analysis and interpretation of patterns in the Pratibha Parv data. Schools were chosen purposively to be outside our sample and include both primary and middle schools in urban and rural areas. In each school, 3 classes were observed over the two days of testing. In total, we visited 52 classrooms across 17 schools. In total 51 classrooms across 17 schools were observed across the two years.

The observations followed a structured protocol which aimed to capture a detailed description of how testing was conducted in Pratibha Parv. This included, for instance, detailed observations about the instructions given to students; whether students were observed copying and what proportion of students were observed doing so; whether the teacher tried to stop cheating from happening; whether the teacher left the classroom during the assessments; whether the teacher was observed trying to help students cheat and, if so, how; whether the external monitor visited the school and, if so, did they observe testing and provide vigilance during the process; details of grading and so on. The intention of the exercise was to, in this small sample, get a comprehensive picture of what testing under business-as-usual looked like. Schools were not informed about our visit before the day it happened. Observation teams arrived with authorization from the state government and further consent for observation was obtained from the school principals before any observations were carried out. Respondents were assured that all observations would be de-identified and the names of individual schools and teachers would not be made public.

Descriptive results

We observed some students cheating from each other in all but one of the 51 classrooms observed. In the vast majority of the cases, over half of the students were observed to be copying at some point during the test. This cheating took multiple forms including copying from each others' answer-sheets and asking each other the answers to specific questions. For the most part, teachers did not attempt to stop this cheating. In several cases, they would admonish students once but then ignore copying as it happened again.

Teachers were observed actively helping students cheat in a substantial minority of classrooms. This included giving answers to individual questions to students, helping them erase and

correct answers and providing hints towards the correct solutions. Note that these are levels in the presence of external monitors and potentially understate the prevalence of such practices significantly.

There was very little evidence of external oversight or monitoring of the assessment process. Although each school is officially assigned to an external official, we did not see such visits happen in most schools over the course of two days of testing. In the schools where the external monitor did visit, this visit was most often perfunctory and only consisted of a brief conversation with the principal and school staff and a look at the paperwork. Only in two cases did we see the external official observe the testing process in detail.

B.2 Qualitative interviews with teachers

We further collected extensive qualitative information based on semi-structured open-ended interviews of school staff and education officials in 6 districts in the 2017-18 academic year. These interviews were broad-ranging and covered, primarily the functioning of the Shaala Siddhi intervention evaluated in Muralidharan and Singh (2019), in addition to substantial discussion of general challenges and constraints faced in the education system including the discussion of Pratibha Parv assessments.

In each of the 6 districts, we randomly sampled three schools: one from the universe of schools assigned to the Shaala Siddhi intervention focused on school management, one more from the a list of "champion schools" which were designated by the government as effective implementers of the program, and one control school (from our sample) for understanding business-as-usual constraints to school effectiveness. In each district, we also randomly sampled one JSK office and one Block Education Office, where we interviewed relevant education officials who are responsible for implementing the program. The aim of the exercise is to provide more context for understanding the failure of the program. Here I restrict my attention to themes that emerge specifically in relation to Pratibha Parv and test-based accountability.

Several themes emerge in the interviews. First, most teachers can explain the rationale behind Pratibha Parv well, many of them report no trouble at all in administering and also provide relatively general responses about the assessments being useful. Yet, several others disagree. Even without being prompted specifically about cheating, some acknowledge the existence of cheating. At least one teacher linked it to the government's priorities (which reflect national policies) of ensuring that students are not in the bottom rung of achievement and, specifically, to the focus of the education system on no grade repetition until Grade 8. Teachers are also skeptical of the efficacy of any anti-cheating measures to counter this. This is borne out in the following (summarized) excerpts from interviews:

"Even in the annual exams, the children are made to copy. What is the use of sending invigilators to different schools and all. On top of it we are told that we should not fail the children till class 8. So even if the students aren't coming to school they are passed. There is one girl who is always absent but confidently says she will pass the exam by copying from others."

"The questions are asked like in a private school. We have to tell the children all the answers to the questions. The question paper is such that the children aren't able to answer. I would rather set my own paper for the children. Because I only know what the children will be able to answer and what they can read."

"Not all students are able to answer the paper. Some students get less marks. I also help them write the answers. I write the answers on the board and tell them to copy it from the board."

While several teachers do say that regular testing helps them assess the achievement levels of the their students, many remain skeptical about the use of these assessments. In particular, they express skepticism for any larger use of the data outside their individual classrooms.

"This is just a formality. There is nothing happening really, it is only on paper. It is far from reality. This is only a waste of money. What are we getting from this? From the entire evaluation that happens throughout the year, setting up a centre head, sending papers from outside and sending them out for correction, there is nothing you are getting from all of this.

Why are they not letting us check our school papers? Are we thieves? What's the guarantee the teachers of the other school correct properly? Some teachers give full marks for an empty paper also. And some of them give less marks for perfect answers. So we might as well check the papers ourselves.

The evaluation should be made only on academic parameters. There should be provision to fail the students also. By just enrolling in class 1, they have completed class 8. If there is a fear of failing, students will definitely start coming to school. Also, if a student fails, they shouldn't get scholarship and uniforms again."

This teacher, quoted above, is the same person who acknowledged overstating results in order to reduce any potential accountability pressures, as quoted in the main text. While these data are not, by themselves, adequate to make any quantitative or causal determination of the prevalence or the motivation of cheating, whether by students or by teachers, they certainly support the view that the assessment system is deeply compromised by such manipulation.

C Clustering analysis of Andhra Pradesh paper and tablet assessments

Using item-level data at the student-level containing students' selected answers for each multiple-choice question in all three observed subjects for the CSF tests, we calculate summary statistics at the grade level for grade 4. Schools in our sample rarely have more than one classroom for a grade. Since only grade 4 was observed in each school, this corresponds to conducting the analysis at the school-level. Before doing so, the item-level data is cleaned as to not contain any students with invalid responses for all items in a subject. Furthermore, we use only the common items in each subject to calculate the components.

The first component is the class mean score in each subject for common items which is equal to the sum of students' scores divided by the number of respondent students in each class. The maximum possible is 24 (the number of common items) and the minimum is 0. The second component is the class mean score standard deviation.

The third component is the class-level homogeneity index, which is equal to the class-mean of the gini indexes calculated for each test item:

$\overline{H_{js}} = \frac{\sum H_{jts}}{T}$

where H is the gini index for item t in subject s and class j, and T is the total number of common test items. The gini index for each test item is calculated as:

$H_{jts} = 1 - \sum(\frac{n_{os}}{N_{js}})^2$

where the ratio $\frac{n_{os}}{N_{js}}$is equal to the ratio of students in class j that has given the answer option o to test item t. This index reaches zero if all students in class j have answered the same option to the test item, implying that there is no heterogeneity for the answers to that item within that class. If all students in that class have answered differently, the gini index reaches its maximum value equal to $\frac{O-1}{O}$, where O is the maximum number of available answer options for the respondents. In our case, each item has five options (A, B, C, D and invalid answer). The maximum value for the gini index is thus 0.8, implying perfect heterogeneity. Equivalently, the class-level heterogeneity index $\overline{H_{js}}$ also ranges between 0 and 0.8 in this case.

Finally, we compute the class level non-response rate for the common items in each subject. It is equal to the sum of invalid responses for all students across all common items in each class, divided by the total number of items answered in each class (i.e. number of items multiplied by the number of students). The non-response rate ranges between 0-1 and is equal to 0 when a class do not have any invalid item responses, and equal to 1 if all responses in a class are invalid.

Using these four statistics, we use principal component analysis to generate two components. These are then used to conduct k-means clustering of the classes, resulting in six clusters

with varying levels of mean scores and standard deviations, non-response rates and response homogeneity. The summary statistics by cluster are shown in the table below with the "extreme" cluster used as the flag for judging potential manipulation highlighted.

Table C.1: Summary statistics of K-means classroom clusters

	(1)	(2)	(3)	(4)	(5)	(6)
Cluster	Score	Standard Deviation	Homogeneity Index	Non-Response Rate	Prop.Paper (%)	Prop.Tablet (%)
Mathematics						
1	10.2	4.9	0.61	0.11	2.24	20.18
2	16.3	4	0.38	0.11	12.76	9.13
3	10.4	4.5	0.54	0.23	2.5	6.21
4	21.2	1.9	0.14	0.02	43.03	4.54
5	17.4	3.8	0.33	0.03	31.45	15.58
6	11.4	4.7	0.56	0.04	8.03	44.36
All clusters	14.5	4	0.43	0.06	100	100
Telugu (Language)						
1	12.5	5.6	0.52	0.19	3.29	9.79
2	18.6	3.3	0.27	0.02	33.16	15.46
3	11.1	4.8	0.6	0.06	3.29	31.28
4	21.3	1.5	0.11	0.01	40.66	5.19
5	14.7	4.5	0.46	0.02	10.79	25.07
6	16.2	4.7	0.4	0.09	8.82	13.19
All clusters	15.8	3.9	0.39	0.05	100	100
English						
1	9	4.3	0.62	0.09	2.37	20.64
2	13.4	4.6	0.48	0.04	14.34	23.63
3	17.7	3.7	0.31	0.03	35	12.23
4	21.2	1.5	0.11	0.01	37.89	2.27
5	13.3	6.3	0.47	0.15	6.84	9.25
6	9.1	2.8	0.57	0.03	3.55	31.98
All clusters	13.6	3.7	0.44	0.05	100	100

Note: The table presents summary statistics for the k-means clusters in each subject. Columns 1-4 present mean values for classroom characteristics. Score refers to the average subject CSF test score. Standard deviation refers to the average score standard deviation. The homogeneity index provides a measure for the answer similarity within classroom, and is equal to zero when all students in a class have given the same answers to all test items. The non-response rate refers to the fraction of invalid answers in a class out of the total number of administered items. It ranges between 0-1 where 0 means no invalid answers were reported and 1 means all reported answers were invalid. Columns 5-6 present the proportion of schools in the treatment arms that exists in each cluster.