Myths of Official Measurement: Limits to Test-Based Education Reforms with Weak Governance*

Abhijeet Singh † Petter Berg ‡

June 13, 2024

Abstract

Assessment-led school reforms are a central pillar of policy packages recommended to address low student achievement in developing countries. We use direct audit evidence to assess the truthfulness of official assessments in a reform that has tested over 6 million students annually since 2011 in the large Indian state of Madhya Pradesh. Comparing responses to the same test questions by the same students shows a doubling of reported achievement in administrative data versus independent tests. This difference is lower, within schools, in grades with multiple test booklets and external grading. Overall, in contexts with weak governance, interventions relying on test-based accountability appear unlikely to succeed without complementary investments to assure data integrity.

*We are grateful to Erich Battistin, Martina Bjorkman Nyqvist, Konrad Burchardi, Luis Crouch, Lee Crawfurd, Jonathan de Quidt, Jishnu Das, Tore Ellingsen, Clement Imbert, Geeta Kingdon, Gaurav Khanna, Karthik Muralidharan, 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 UK Aid. We are grateful for support to Ms. Deepti Gaur Mukherjee, Mr. Lokesh Jatav and Mr. K.P.S. Tomar from the Government of Madhya Pradesh. Ramamurthy Sripada, Urmi Bhattacharya, Ghazal Gulati, Krishanu Chakraborty, Nawar Al Ebadi, Aditi Bhowmick, Sabareesh Ramachandran, Akankshita Dey and Edoardo Bollati provided outstanding field management and research assistance.

†Stockholm School of Economics; J-PAL; CES-Ifo. E-mail: abhijeet.singh@hhs.se

‡Stockholm School of Economics. E-mail: petter.berg@phdstudent.hhs.se

1 Introduction

Learning levels in many low-and-middle-income countries (LMICs) are abysmally low, a situation widely referred to as a 'global learning crisis' by development agencies (UNESCO, 2013; World Bank, 2018). Two prominent explanations for this status quo are that (i) education systems in LMICs have, until recently, often prioritized improving access but not learning (Pritchett, 2013), and (ii) accountability of schools and teachers is very low in many settings (Mbiti, 2016; Muralidharan et al., 2017). Although experimental evidence suggests that improving accountability can substantially boost learning outcomes (see Glewwe and Muralidharan (2016)), achieving system-wide adoption of these interventions or demonstrating similar effects at scale has proven very difficult (Evans and Yuan, 2020).

Policy proposals to address the learning crisis emphasise a system-wide approach to reform rather than stand-alone interventions from the experimental literature. For instance, the World Development Report 2018 recommends three sequential policy actions to solve the learning crisis: "(i) Assess learning to make it a serious goal, (ii) Act on evidence, to make schools work for learners and (iii) Align actors, to make the system work for learning" (World Bank, 2018). Central to this agenda is the assumption that large-scale student assessments, if adopted by governments, can build consensus about the severity of the learning crisis and serve as a sufficiently reliable basis for targeting and policy action. Assessment-led reforms have a long history, and substantial evidence, in OECD countries like the US (see e.g. Figlio and Loeb (2011)). However, it remains an open question whether countries with weaker state capacity can similarly maintain *reliable* assessment systems over time (Rossiter et al., 2023).

We study this question in the context of a large standardized state assessment that tests all students enrolled in government primary or middle schools in a large Indian state, Madhya Pradesh, annually (∼6-7 million students, across 110,000 schools). This assessment, called *Pratibha Parv*, generates test scores providing the only comparable measure of student and school performance in the state. These scores do not carry formal incentives for teachers or students but are meant to prioritize students and schools for remedial action. The program was designated as a "national best practice" (NITI Aayog, 2016), and also serves as a template for national policy: the new National Education Policy mandates similar tests in key grades across all state education systems (Government of India, 2020). In this paper, we investigate whether these tests reliably measure true student performance. This is central for evaluating this particular reform but also — since both the intervention and (weak) state capacity are typical of many LMICs — informative about the potential of such reforms in similar contexts.

Our empirical analysis focuses on the truthfulness of reporting in these official tests. We are concerned, in particular, about misreporting of achievement levels due to student copying, grade manipulation by teachers, or teachers assisting students in answering exam questions. We take a direct, audit-based approach to detecting such manipulation. Specifically, we compare students' reported responses in the official test in January 2017 with their responses to the *same* test questions in a retest, administered a month later, in a representative sample of 283 primary and middle schools. This retest was independently proctored and graded to deter cheating and serves as a manipulation-free external benchmark.

We use this audit 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 — a doubling of reported achievement. This discrepancy between the proportion of correct responses across the two assessments is present across the full distribution of prior achievement but substantially larger for weaker students.

Second, this discrepancy in achievement levels reflects active manipulation and is not merely an artefact of, say, a difference in student effort across the two tests or a fade-out of short-term test preparation. The magnitude of the discrepancy is reduced by one-half in Mathematics, and by three-quarters in Hindi, in middle school classes that mandated multiple booklets and external grading to deter cheating (compared to other classes in the same school). Direct observation of testing in a subset of schools further confirms both student copying and teacher-assisted manipulation of scores. These official assessments thus mislead inferences about learning levels: schools report, on average, only ∼8% of students failing in grade-level tests, even though fewer than one-third of Grade 5 students in this state can read a Grade 2 level text (Pratham, 2017).

Third, although the achievement levels in official data are substantially exaggerated, we find that the ordinal ranking of schools, based on the school-level aggregate score, is remarkably stable over time. This is despite aggregate scores showing dramatic (implausible) improvements over time and a particularly sharp reduction in the number of students and schools classified as failing. Ordinal ranks, both at school level and for individual students, are also strongly correlated across the official tests and our independent audit. This manipulation of reported *levels* of achievement, but not ordinal ranks, likely reflects the emphasis of the assessment on absolute achievement rather than relative comparisons.

Our principal contribution is to provide the most comprehensive evaluation of the promise of standardized assessments in LMICs to address learning deficits. Few studies in non-OECD

contexts test the truthfulness of large official assessments, and none do so with direct retest-based audits.[1]

We provide three insights to a large literature on how to improve accountability and student achievement in LMICs (Mbiti, 2016; Ganimian and Murnane, 2016). First, our findings suggest large system-wide assessments will likely be inadequate as a barometer of learning levels, or as a basis for policy action, given the weak institutional governance of education systems in South Asia and sub-Saharan Africa. This is in marked contrast to the widespread use of these measures in upper-middle-income countries both to measure learning and as a basis for policy.[2] Second, the prevalence and scale of test score manipulation here indicate a substantial barrier to scaling up many interventions that have been shown to be effective in multiple experiments in similar contexts: this includes, for example, providing student and school report cards to parents (Andrabi et al., 2017; Afridi et al., 2018) or performance related pay that depends on test score levels or gains (see, e.g., Muralidharan and Sundararaman (2011); Loyalka et al. (2019); Mbiti et al. (2019); Leaver et al. (2021)). System-wide adoption of these interventions will require complementary investments for ensuring data integrity (e.g. Singh (2024)). Third, the widespread adoption of large scale assessments in India, despite local knowledge of both manipulation and ineffectiveness, provides a particularly stark example of "isomorphic mimicry", the practice of copying reforms seen as "best practice" despite ineffectiveness (DiMaggio and Powell, 1983; Pritchett, 2013).

More broadly, the dependence of policy measures on underlying official measurement is relevant beyond education: the measurement of service delivery outcomes underpins reforms focused on incentives, accountability or information in many sectors.[3] In this respect, official data form a core pillar of state capacity (Scott, 1998). Thus, our results also relate directly to studies documenting misreporting in administrative data in LMICs with independent measurement (see, especially, Reinikka and Svensson (2004), Olken (2007), Duflo et al. (2013)

[1]The only exception is Singh (2024) featuring a complementary study, designed subsequent to results reported here, to reduce manipulation through digital testing in a different state (Andhra Pradesh). Importantly, that experiment was confined to a single district and focuses on a one-off official test rather than a longstanding assessment. See also Berkhout et al. (2024) who do not directly measure cheating but attempt to study a reform to improve data integrity.

[2]In the United States, for example, a substantial literature evaluates similar test-based accountability mechanisms (see, e.g., Rockoff and Turner (2010); Figlio and Loeb (2011); Reback et al. (2014). Similarly, large-scale system-wide assessments have long been the basis for both educational policy and research in Europe and much of Latin America (see, e.g., Mizala et al. (2007) and Mizala and Urquiola (2013) in Chile).

[3]For instance, the World Bank reports 'results-based financing', such as development impact bonds, in many sectors including energy, education, healthcare, 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.

and Niehaus and Sukhtankar (2013)). In contrast to this past work, we demonstrate that misreporting may be stark even without direct financial incentives, highlighting that the threat to administrative data integrity is much broader than financial corruption alone. Improving the underlying structure of administrative data may, therefore, be an important challenge not just for research but also for improving state capacity for development.

2 Background

2.1 Context and policy design

Our study is set in Madhya Pradesh, 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 state of public education in Madhya Pradesh serves as a stark illustration of the fundamental challenges facing India's elementary school system. In 2016, only 31% of Grade 5 students in government school were able to read a text at Grade 2 level, and 15% could perform division (Pratham, 2017). Rates of teacher absence are low with 26% of teachers absent during school visits (Muralidharan et al., 2017).

The testing regime we study, called *Pratibha Parv*, was launched in 2011 by the Government of Madhya Pradesh, an early adopter of large-scale assessments, to address the poor quality of public education. Under this program, a standardized grade-specific assessment is administered annually to all students in Grades 1-8 of the public schooling system (which includes about 110,000 schools and nearly 7 million students aged 6 to 14). The program has been designated a national best practice and, with the National Education Policy of 2020, similar assessments are now planned nationally (NITI Aayog, 2016; Government of India, 2020).

These assessments, based on policy documents, 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. Importantly, the assessment is intended to be a diagnostic exercise and does not formally have high stakes. For example, 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).[4] However, as we discuss in Section

[4]This setting, thus, contrasts with the high-stakes environments in many previous papers featuring explicit monetary incentives (e.g Martinelli et al. 2018) or where the test results had significant long-term

4 with supporting qualitative evidence, teachers may still experience implicit pressures to display good test results to avoid potential scrutiny.

The assessments are held in December or January, which is in middle of the school year. Students are tested in several subjects, with questions reflecting the school curriculum from the previous and current grade (i.e. material that should have been taught in the past 12 months). The test includes both multiple-choice questions and open-ended response questions.[5] In the first two years of primary school, the assessment includes questions requiring an oral response (that are meant to be administered individually by the teacher) but exclusively comprises of written responses from Grades 3–8.

Tests are proctored and graded by teachers in the same school with, de facto, little external oversight of test administration. 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. Reports of such cheating are common, including in the local media.

2.2 Data

The audit study in Madhya Pradesh was embedded in a larger data collection (Muralidharan and Singh, 2020). From 2016 to 2018, we collected primary data on student achievement in a panel of 298 schools, spread across 10 districts in the state (Figure A.1).[6] We conducted independent student tests in math and Hindi (language) in these schools in July 2016, February 2017 and February 2018. These assessments were independently-proctored by surveyors hired by the research team and graded centrally. Further, they were designed to capture a broad range of learning outcomes and, in each round, test papers were the same within each grade but varied across grades. A subset of test items were taken from the *Pratibha Parv* assessments.[7]

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 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)

[5]The proportion of multiple-choice questions (MCQ) has varied over time. In 2017, the year of the audit study, the test comprised entirely of MCQ items in Grades 3, 5 and 8 but all grades had some MCQ items.

[6]These districts were Bhopal, Raisen, Rajgarh, Sehore and Vidisha in the Bhopal region and Alirajpur, Barwani, Dhar, Jhabua and Khargone in the Indore region, where we focused especially on officially-designated "tribal blocks". The cumulative population of these districts was about 15 million people in 2011 and they represent the social and economic diversity of the state, which has 51 districts in total.

[7]The principal aim of the independent test administration was to provide the primary outcomes for the evaluation in Muralidharan and Singh (2020). The common items allow us to use these as direct audits.

In surveyed schools, we transcribed student level scores in each subject from the official *Pratibha Parv* assessments preceding each round of our independent tests. In the 2016-17 academic year only, when schools were instructed by the government to also record item-level data, we additionally transcribed data on student responses to individual test items. This data was available in 283 schools, which provide our main analysis sample. Table A.1 compares these sample schools to the full population of schools in the state and in our study districts. Observed characteristics, including the distribution of test scores, are similar across schools in the survey sample, the population of schools in these districts, and in the state overall.

The audit compares item-level responses from the *Pratibha Parv* assessments in January 2017 to the independent testing in February 2017. To avoid ambiguity in grading, we will restrict attention to only those questions which were multiple-choice in the official assessment (and therefore in our retest) and admit only one correct unambiguous response.

We match students across tests based on their "scholar number", which is unique within the school over time, and student names. The main reason for not being able to match students across the two assessments is student attendance: while nearly all students are present for official tests, approximately 53% of students were present on the day of the independent test.[8] Matched item-level data is available for 4676 students from Grades 3 to 8, which is reduced to 3435 students when restricting to multiple-choice questions only.[9] Independent tests were not pre-announced, so absence on the day of the test does not represent self-selection. Figure A.2 shows the distribution of achievement on the official test for matched students compared to the full population. These distributions are very similar and our analysis sample represents the full distribution of achievement. There is a modest positive selection into the retest (as expected given a positive correlation between attendance and achievement), suggesting that our estimates may modestly understate the magnitude of test distortion.

2.3 Descriptive statistics

The evolution of *Pratibha Parv* school scores between 2012-13 and 2016-17 already provides suggestive evidence of manipulation (Table 1). In this period, average *reported* achievement in the assessments rose sharply from 57 points to 69 points, out of 100 (roughly a one standard deviation increase). Dispersion reduced sharply as well. Finally, while 21% of schools obtained "failing" grades ("D" or "E") in 2012-13, this number dropped to 3% in

[8]This is similar to student attendance rates of 54.8% in middle schools and 58.5% in middle schools documented for representative samples at the state level by the ASER Report in 2016 (Pratham 2017).

[9]This drop in observations is mainly caused because we had not included common MCQ items in Grade 4. This causes Grade 4 students to be dropped from the sample.

2016-17 with a sharp corresponding reduction in the proportion of failing students. This near-elimination of poor performance is not matched in independent data.[10]

3 Misreporting in official test data

In this section we present comparisons of student performance in the official *Pratibha Parv* assessment and independently audited retests, and link it to active manipulation.

3.1 Direct audit comparisons

Figure 1 presents the main result of the audit. Each dot in the plot represents an individual multiple-choice test question. The proportion of students reported to answer correctly is shown on the horizontal axis for the independent audit and on the vertical axis for the official test. With the exception of a single question in mathematics, a much larger proportion of students is reported to answer the same item correctly in the official data than in independent retest. The magnitude of this discrepancy seems to be higher in mathematics than in Hindi. Aggregated at the student level, the proportion of correct responses to the same multiple-choice questions is, on average, 38.9 percentage points higher in math and 33.8 percentage points 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.

Next, in Figure 2, we investigate if the discrepancy varies in size across the achievement distribution. Discrepancy is measured as the difference in the proportion of correct responses in the official test and the retest; it is bounded between 1 and -1.[11] We use the official test scores in the previous school year as the measure of lagged achievement.[12] The left panel shows non-parametric plots of the relationship between discrepancy, aggregated at the student level, and the student's percentile in the overall score distribution in that subject/class in the previous year. In the right panel, we plot the conditional expectation of the student's score on the two later assessments by their position in the baseline achievement distribution in December 2015; here, the expected discrepancy is given by the distance between the two curves. In both subjects, discrepancy is positive across the full distribution but substantially larger for weaker students. In Hindi, the discrepancy is just above 40

[10] ASER data report that, 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 figures stood at 10.2% and 8.4% (Pratham, 2017).

[11] A value of 1 indicates that the student answered all items correctly in the official test, but none in the retest, while -1 indicates the opposite case.

[12] We use official scores as a measure of lagged achievement as this is available for most students who were enrolled in the same school in the previous year. Results are substantively similar if, instead, we use independent tests administered in July 2016. The independent tests have the virtue of being externally proctored and graded but exclude students who were absent on the day of the test.

percentage points in the bottom decile and under 15 percentage points in the top. The decline is less stark in math (with higher discrepancy): moving from the bottom to top decile, the discrepancy reduces from 45 to 30 percentage points.[13]

3.2 Discrepancy between assessments reflects cheating

Although Figures 1 and 2 strongly suggest cheating, these patterns are also consistent with other explanations. If students exert lower effort in the independent assessments, perceiving them to have lower stakes than official tests, that could in principle explain the discrepancy that we find. Further, while official tests are scheduled and pre-announced, the independent assessments are not. If anticipated exams are preceded by extensive preparation, as is common in many education systems including in India, the discrepancy could merely reflect this "test preparation effect". We take two distinct approaches to document that these explanations are unlikely to account for this discrepancy.

The first is direct observation. In 2016-17 and 2017-18, we conducted observations of the test administration in 51 classrooms across 17 schools in 4 districts. These observations were authorized by the state government but schools were not informed of our visits in advance.[14] In 50 (out of 51) classrooms, we found some form of student copying. Most teachers did not actively try to control such copying and in 20 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 students and to help them erase and correct the answer, if the student had answered incorrectly. External monitors, while formally appointed for each school, were rarely present in classrooms. If they did come to the school, they stayed only a short while and did not directly affect test administration. This pervasiveness of cheating even in the presence of external surveyors, which is also indicated by reports in local media, suggests that the discrepancy between official and audited assessments that we find reflects cheating.

Second, we adopt an alternative approach which exploits variation in anti-cheating measures across grades to evaluate whether a reduction in the opportunity for students and teachers

[13]One concern in the interpretation above may be mean reversion in test scores due to, for example, measurement error in baseline achievement (Chay et al., 2005; Jacob and Rothstein, 2016). Note though that the discrepancy is measured here as the difference between two later assessments, *both* of which should be affected by mean reversion. Our estimates will only be confounded if there is *differential* mean reversion, from Dec 2015, between tests administered in January 2017 and February 2017, at a given percentile. This seems unlikely.

[14]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 purposely selected to cover multiple districts and types of schools. A detailed description of this exercise is provided in Appendix B.

to cheat in the official assessments translates into lower discrepancy with audit results. In January 2017, the official assessments introduced multiple test booklets in Grade 8 and also mandated that answer scripts to be sent to a different school for grading.[15] If differences in student effort between tests, or the rapid fade-out of short-term test preparation, do not differ across grades, any reduction in the discrepancy for Grade 8 would reflect the effect of multiple booklets and external grading. Panel A of Table 2 reports the results of this investigation. Student-level average test scores, on the same multiple-choice questions, are higher by $\sim$ 54 percentage points in Math in Grade 7 in the official assessment than the retest, but this discrepancy is reduced by half in Grade 8 (Col. 1); conditioning on school fixed effects does not affect these estimates (Col. 2). In Hindi, the discrepancy is lower at about 36 pp in Grades 6 and 7, but is reduced by 27 percentage points in Grade 8 (Cols. 3 and 4). Reported coefficients are all significant at the 1% level.

We can investigate this further using the fact that each pair of responses by a student to the same item, across the two assessments, can be treated as a single audit. Specifically, consider the following specification:

$$Y_{qig2} = \alpha_q + \beta_1.(T_g \times Y_{qig1}) + \beta_2.Y_{qig1} + \epsilon_{qig2} \tag{1}$$

where Y_{qigr} is an indicator variable for a correct response to an individual question q, answered by student i, in grade g at testing round r (where 2 denotes our retest in February and 1 the official test in January). T_g is an indicator variable for being in Grade 8 ("treated"). In this specification, β_2 captures the extent to which having answered a question correctly in the official assessment increases the predicted probability of also answering it correctly in the independent assessment; β_1 captures whether this correspondence between the official and independent assessments is higher for treated classes. A consequence of reduced distortion should be to increase the correlation between answering the same item correctly in the official test and the retest.[16]

Results are reported in Panel B of Table 2. Answering correctly in the official test is associated with a 4.3 percentage points higher probability of answering correctly in the audit in mathematics, and 13.2 percentage points in Hindi, in untreated grades (Cols. 1 and 3). In Grade 8, however, this probability is higher by roughly 9 percentage points in math and 7.5

[15] Both these measures are commonly employed to deter copying and grade manipulation in India, for example, in high-stakes examinations at the end of secondary school.

[16] The main assumption is that there are no other systematic differences between the grade-specific tests in measurement error or persistence of true ability over one month. This assumption is plausible in this setting since we only compare multiple-choice items with unambiguous answer choices. Item-specific fixed effects control for any mean differences in difficulty across test questions and also absorb grade differences. Thus, differences in test content should not lead to bias.

percentage points in Hindi with the difference being statistically significant at the 5% level in math and the 10% level in Hindi. These figures are conditional on item-level fixed effects, i.e. they control for question-specific characteristics that raise/lower the proportion of correct responses for the item in our independent assessments. Adding school fixed effects does not affect the magnitudes appreciably. Official test scores being more predictive of future outcomes in settings with higher barriers to cheating, is consistent with the discrepancy between the assessments reflecting cheating (which is reduced in Grade 8).

3.3 Preservation of ordinal information

A distinct empirical question is whether official data still contain useful information about the ranking of schools and students. We investigate this in Figure 3.

Panel A shows a substantial positive association between the percentile ranks of schools and students between the official assessment and the audit. This is true both for the ranking of schools and individual students, in both subjects. Thus, despite severe misreporting of the levels of absolute achievement, it appears that even the distorted metric preserves substantial information about ordinal ranks.

Panel B further shows that the ranking of individual schools has substantial persistence across years. Although there was a sharp decline in the proportion of low-performing schools (Table 1), the percentile ranking of schools appears remarkably stable.

In principle, the preservation and stability of this ordinal information allows for instituting several policy measures. We discuss this possibility in Section 4.2.

4 Discussion

4.1 Why is cheating widespread?

A final set of concerns relate to the motivations for why individuals cheat in this setting, even without formal incentives, and why governments continue to scale these tests up. These questions are not the principal focus of our analysis and we cannot provide definitive evidence. However, we suggest potential explanations.

That teachers cheat, whether actively by inflating grades and assisting students or passively by letting them copy from each other, may be rationalizable with the incentives they face as civil service workers. Even though their pay and tenure do not depend on the outcomes of the test, they may still face pressure from higher officials in the education system if their schools are seen to be "failing". We conducted qualitative interviews to supplement our understanding of the education system and the incentives faced by teachers and education

officials which provides some evidence consistent with this explanation. As one teacher reported:

> 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 Shiksha 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.

Administering assessments, grading and reporting truthfully requires teachers to expend effort, accurate reports of (low) learning levels may incur career costs, while providing inflated measures of achievement carries no penalties; in combination, these could account for the pervasiveness of over-reporting.[17]

Why do students cheat on exams that do not determine grade progression or certification that affects longer-term outcomes? One plausible explanation is that students, already in primary schools, have internalized the high-stakes consequences of test scores (which now only apply in later years) and the overall orientation of the school system which emphasizes the centrality of school test scores as the only relevant outcome.

Finally, why do governments scale these up? In common with bureaucratic incentives in many settings, the education system here rewards administrative compliance, the appearance of activity and judges programmatic success by paperwork completed. By those metrics, and regardless of true outcomes, the *Pratibha Parv* is a resounding success (and indeed recognized by the national government as such).[18] That governments across India persist with administering and scaling up these tests may, in turn, be explained by organizational

[17] See, e.g. Angrist et al. (2017), who cite shirking from the effort of transcribing results honestly (which is cumbersome) for why teachers in Southern Italian provinces inflate scores.

[18] These features of the education system in M.P., and the extent to which it replicates bureaucratic incentives in other sectors, is discussed in greater detail in Muralidharan and Singh (2020). That discussion relates to a large school management reform that was successful on paper and scaled up nationally, despite experimental evidence that it had not improved any outcomes — the underlying bureaucratic structure and incentives are identical across policies, including the assessments regime we evaluate here.

incentives for "institutional isomorphism" wherein various practices take on "a ritual aspect; [organizations] adopt these "innovations" to enhance their legitimacy, to demonstrate they are at least trying to improve." (DiMaggio and Powell, 1983).

4.2 Implications for policy and research

Our results suggest that state-led assessments, as currently administered in India, severely inflate official measures of student learning. They understate the 'learning crisis' and, in doing so, fail to even diagnose the problem they are meant to fix. Expectations by national and state governments, and multilateral agencies such as the World Bank, that these assessments will serve to catalyze action to focus on low achievement seem optimistic. In the Indian case, by delaying government recognition of low learning levels (and consequently, policy action), distorted official metrics may have been worse than no official metrics at all.

That official assessments still contain some ordinal information could, in principle, suffice for many policy uses: this includes performance-based pay (Barlevy and Neal (2012)), the generation of school league tables and reforms that target relatively low-performing schools and students. Unfortunately, even this seems optimistic. A plausible reason for why levels are severely distorted but ordinal ranks appear not to be is that all administrative attention in the program was related to levels/grades and not the ranking of schools.[19] If policies emphasized relative ranks instead, it is plausible that such a switch would distort the ordinal information as well.

We interpret our results as suggesting that any use of these test score data for downstream interventions will likely require simultaneous enabling reforms to ensure data integrity. The costs and feasibility of these enabling reforms need to be explicitly accounted for before deriving policy implications from prospective evaluations. For example, meta-analyses of cost-effectiveness of educational interventions, such as presented by Kremer et al. (2013), which are intended 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 about the prospect of using such data for research. Not only is the level of achievement distorted, there is suggestive evidence that such manipulation may differ over time and varies across students over the achievement distribution (Figure 2). Thus, unlike past experience in the higher-income OECD settings where administrative data have revolutionized the scope of education research (Figlio et al., 2016), even low-stakes

[19]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)).

assessments are unlikely to be reliable bases for quantitative research. The returns from improving the integrity of these data systems are potentially very large (see Singh (2024), for example, for experimental evidence on digital testing).

5 Conclusions

This paper evaluated the potential of assessment-led education reform to diagnose the state of the learning crisis in LMICs, and provide a reliable ground for policy action. Specifically, we showed that a major assessment-led reform in a large Indian state severely exaggerates true learning levels across the full distribution of assessed children, and particularly so at the bottom. These results reflect steady-state implementation measured 5-6 years after program introduction and not merely teething problems. Many Indian states have instituted similar tests in recent years, and with the National Education Policy 2020, similar assessments will be scaled up nationally across the public schooling system. Overall, our results cast severe doubt on the effectiveness of these measures, as currently administered, in achieving its aim of providing accurate information on academic achievement to policymakers. Weak state capacity and low learning levels are common across education systems in many LMICs; our results are likely to be informative also in these settings.

More fundamentally, our results highlight the lack of *reliable* administrative information as both the outcome of poor state capacity and a constraint for improving public sector governance. While education is an important sector in its own right, similar problems of data integrity span sectors and affect public service delivery more broadly. While this is sometimes the result of financial corruption (see, e.g., Niehaus and Sukhtankar (2013)), we show it can also emerge in settings without financial or formal incentives. For outcome-based decision-making or incentives to be adopted in the public sector in LMICs, addressing data integrity appears to be central. Research on how to achieve this at scale is likely to be informative and of independent interest across contexts.

References

Afridi, F., Barooah, B., and Somanathan, R. (2018). Improving learning outcomes through information provision: Experimental evidence from Indian villages. *Journal of Development Economics*.

Andrabi, T., Das, J., and Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, 107(6):1535–63.

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

Barlevy, G. and Neal, D. (2012). Pay for percentile. *American Economic Review*, 102(5):1805–31.

Berkhout, E., Pradhan, M., Suryadarma, D., Swarnata, A., et al. (2024). Using technology to prevent fraud in high stakes national school examinations: Evidence from Indonesia. *Journal of Development Economics*, 170:103307.

Bertoni, M., Brunello, G., and Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics*, 104:65–77.

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

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

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

Dee, T., Dobbie, W., Jacob, B. A., and Rockoff, J. E. (2019). The causes and consequences of test score manipulation: Evidence from the New York regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. *NBER Working Paper*, (w22207).

DiMaggio, P. J. and Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, pages 147–160.

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

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

Figlio, D., Karbownik, K., and Salvanes, K. G. (2016). Education research and administrative data. In *Handbook of the Economics of Education*, volume 5, pages 75–138. Elsevier.

Figlio, D. and Loeb, S. (2011). School accountability. In *Handbook of the Economics of Education*, volume 3, pages 383–421. Elsevier.

Ganimian, A. J. and Murnane, R. J. (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research*, 86(3):719–755.

Glewwe, P. and Muralidharan, K. (2016). Improving education outcomes in developing countries – Evidence, Knowledge Gaps, and Policy Implications. *Handbook of the Economics of Education*, 5.

Goodhart, C. A. (1984). Problems of monetary management: the UK experience. In *Monetary Theory and Practice*, pages 91–121. Springer.

Government of India (2020). *National Education Policy 2020*. Ministry of Human Resource Development, Government of India, New Delhi.

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

Kremer, M., Brannen, C., and Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, 340(6130):297–300.

Leaver, C., Ozier, O., Serneels, P., and Zeitlin, A. (2021). Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools. *American economic review*, 111(7):2213–2246.

Loyalka, P., Sylvia, S., Liu, C., Chu, J., and Shi, Y. (2019). Pay by design: Teacher performance pay design and the distribution of student achievement. *Journal of Labor Economics*, 37(3):621–662.

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., and Rodrigo, R. (2018). Cheating and incentives: Learning from a policy experiment. *American Economic Journal: Economic Policy*, 10(1):298–325.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., and Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. *The Quarterly Journal of Economics*, 134(3):1627–1673.

Mbiti, I. M. (2016). The need for accountability in education in developing countries. *Journal of Economic Perspectives*, 30(3):109–32.

Mizala, A., Romaguera, P., and Urquiola, M. (2007). Socioeconomic status or noise? Tradeoffs in the generation of school quality information. *Journal of development economics*, 84(1):61–75.

Mizala, A. and Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics*, 103:313–335.

Muralidharan, K., Das, J., Holla, A., and Mohpal, A. (2017). The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics*, 145:116–135.

Muralidharan, K. and Singh, A. (2020). Improving public sector management at scale? Experimental evidence on school governance India. Working Paper 28129, National Bureau of Economic Research.

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

Niehaus, P. and Sukhtankar, S. (2013). The marginal rate of corruption in public programs: Evidence from India. *Journal of Public Economics*, 104:52–64.

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

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

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

Pritchett, L. (2013). *The Rebirth of Education: Schooling Ain't Learning*. Brookings Institution Press.

Reback, R., Rockoff, J., and Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, 6(3):207–41.

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

Rockoff, J. and Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy*, 2(4):119–47.

Rossiter, J., Abreh, M. K., Ali, A., and Sandefur, J. (2023). The high stakes of bad exams. *Journal of Human Resources*.

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

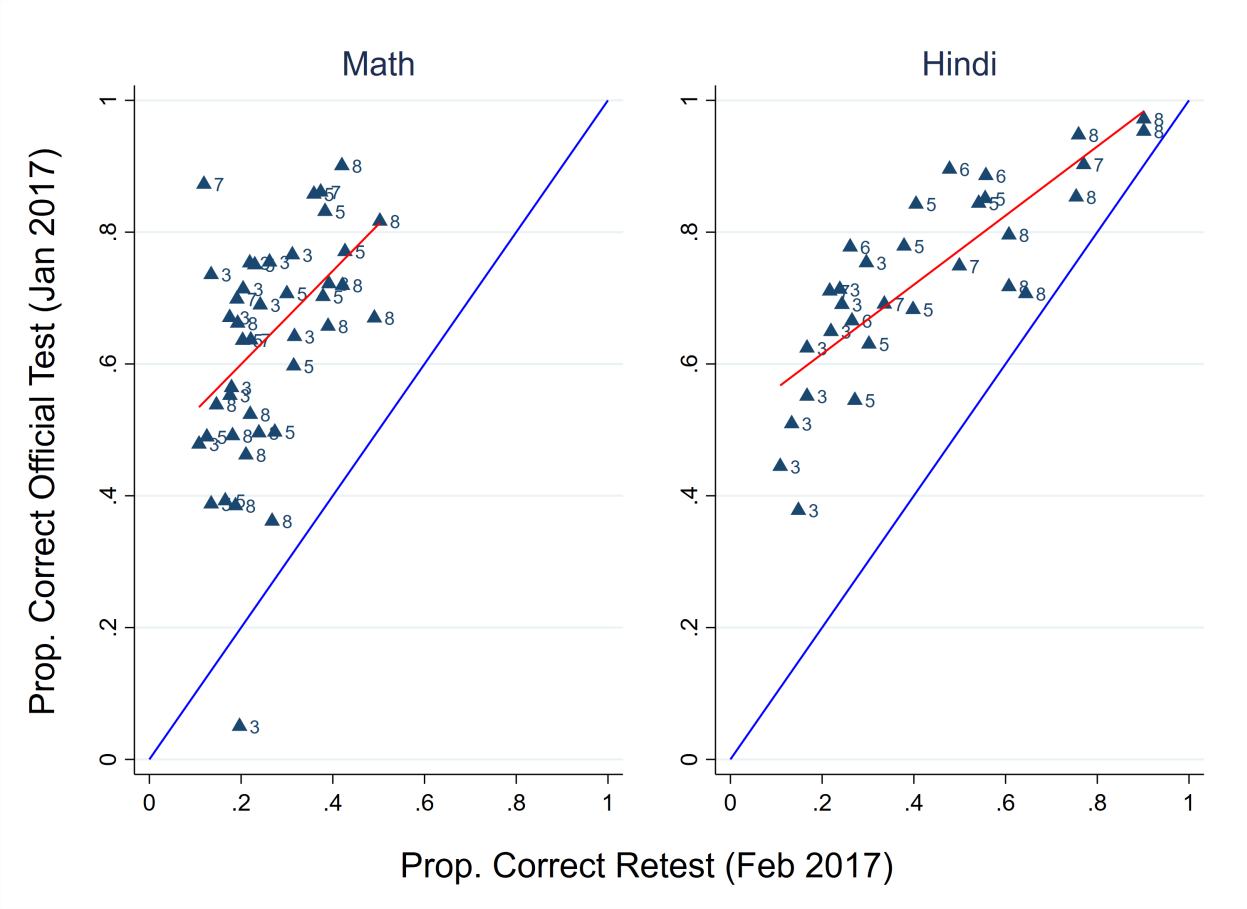
Singh, A. (2024). Improving administrative data at scale: Experimental evidence on digital testing in Indian schools. *Economic Journal*, Forthcoming.

UNESCO (2013). *The Global Learning Crisis: Why every child deserves a quality education*. UNESCO, Paris, France.

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

6 Figures

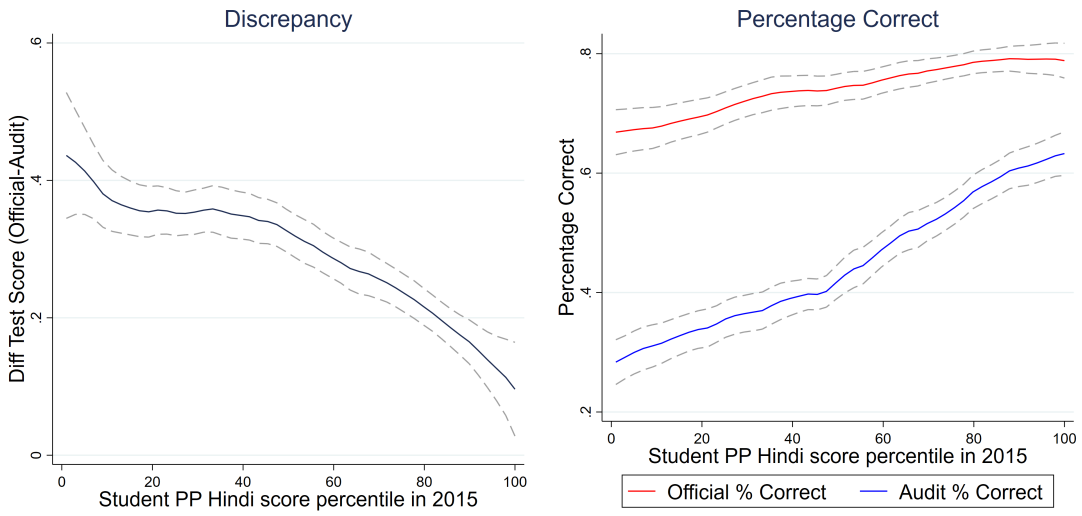
Figure 1: Comparing item-level data from official tests and retest audit



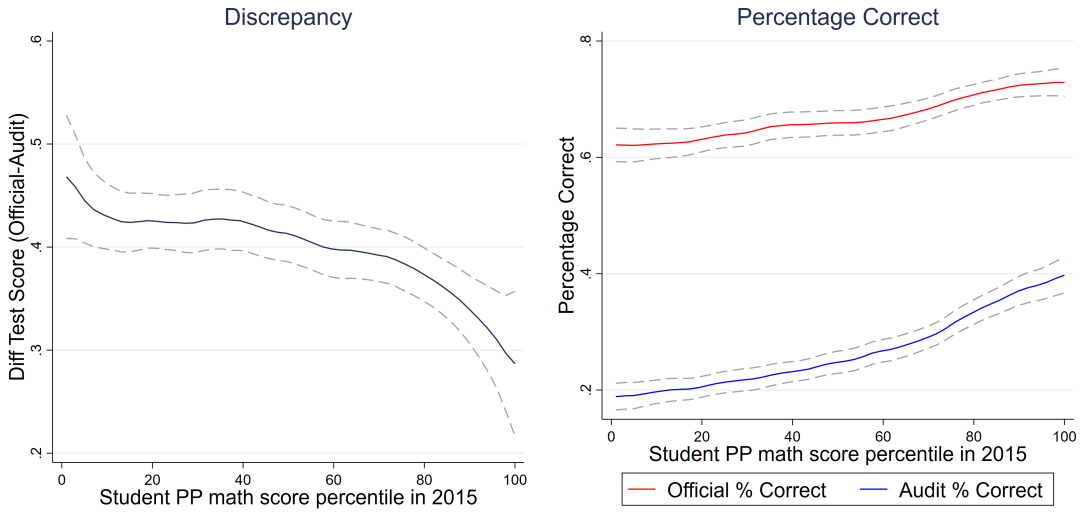
Notes: Each dot in this figure is an individual multiple-choice test question and compares the proportion of students who are reported to have correctly answered in the Pratibha Parv assessment (Jan 2017) with the percentage correctly answered in the audit (Feb 2017). There are 69 such test questions across the two subjects. The marker label indicates the grade in which the question was administered.

Figure 2: Discrepancy over the achievement distribution

(a) Hindi



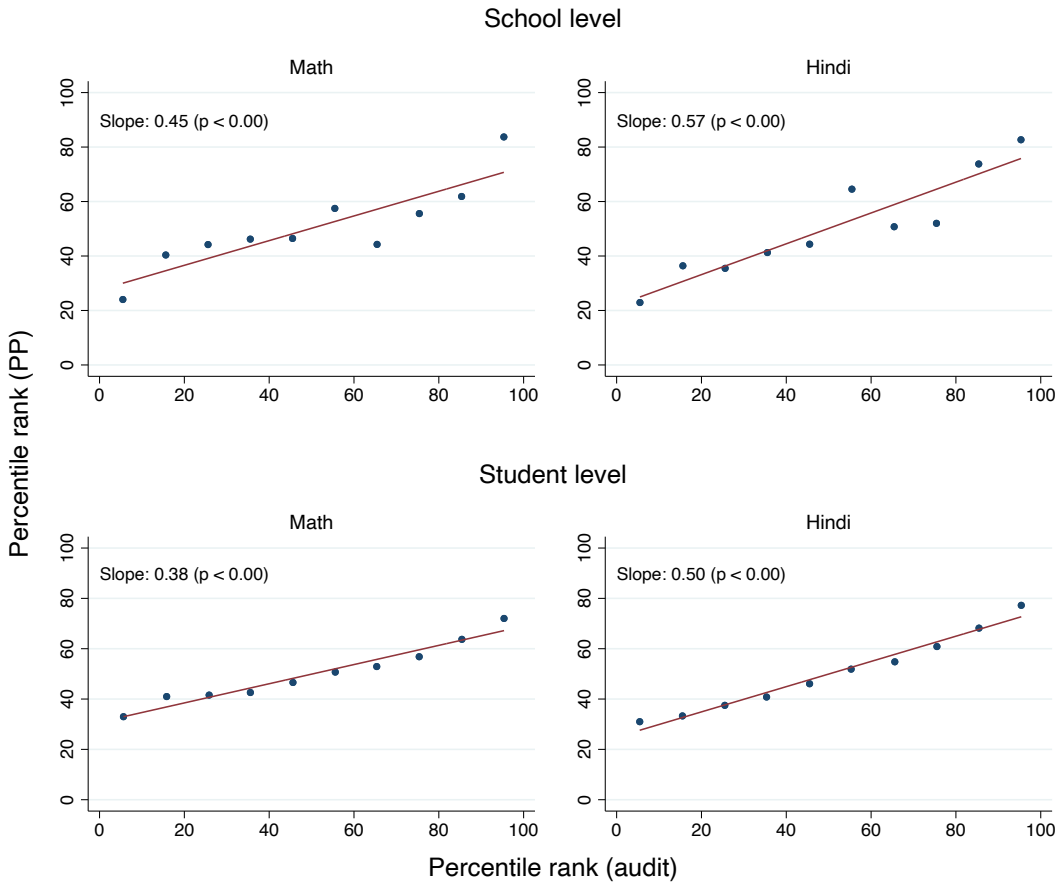
(b) Math



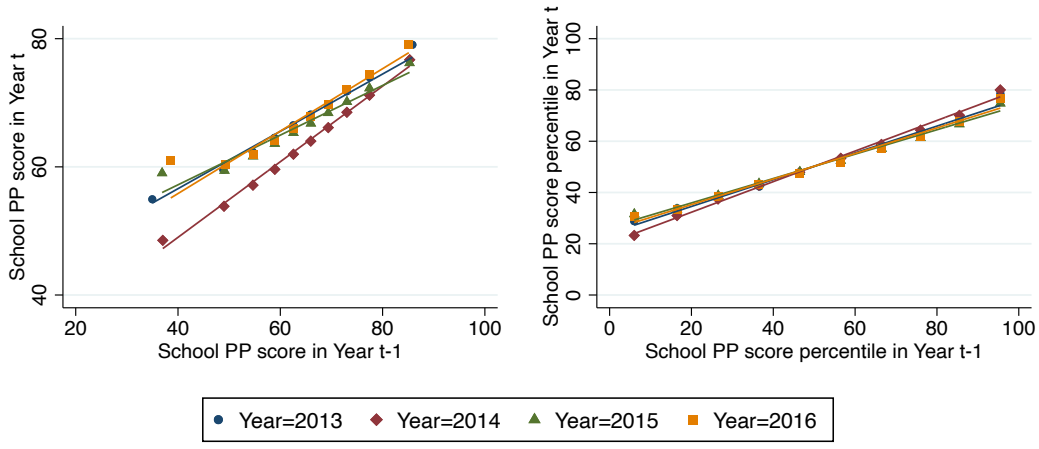
Notes: Discrepancy is defined as the difference between proportion correctly answered in the official assessment and the retest for test items which are common across both assessments. Percentiles are defined over the test score in the previous academic year. The left panel shows the variation of discrepancy 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 across the same percentiles. The distance between the two curves provides the discrepancy measure. This analysis does not include 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 since lagged achievement is not available for them. The conditional means are estimated using local polynomial smoothing with a bandwidth of 10 using the Epanechnikov kernel, shown with 95% confidence intervals. The number of underlying observations is 1463 in Hindi and 1762 in math.

Figure 3: Ordinal ranks of schools and students in Pratibha Parv

(a) Correlation of ordinal ranks between official and audit scores



(b) Stability of school ranks in official scores across years



Notes: In Panel A, we relate the percentile rank of schools and students in the PP score distribution (vertical axis) and the corresponding percentile rank in the audit score distribution (horizontal axis). The sample contains 7248 students and 291 unique schools. In Panel B, we relate absolute scores at the school level (left panel) in the official data, and the percentile in the within-year distribution (right panel). The sample consists of 114,398 unique schools. Averages in 10 equally sized bins are shown with linear fitted lines.

7 Tables

Table 1: Distributions of school scores by year

	2012	2013	2014	2015	2016	Pooled
School Pratibha Parv score	57.04	64.68	64.01	66.98	68.98	64.36
	(15.52)	(13.48)	(13.44)	(11.36)	(11.63)	(13.77)
Enrolment	69.84	83.02	75.97	63.20	63.09	71.05
	(58.33)	(73.73)	(59.81)	(52.47)	(51.06)	(60.14)
% of schools with grade A	0.12	0.23	0.21	0.25	0.31	0.22
% of schools with grade B	0.32	0.44	0.44	0.50	0.49	0.44
% of schools with grade C	0.35	0.26	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	0.07	0.02	0.02	0.01	0.01	0.02
Number of schools	109842	112408	112869	112346	111138	558603

Notes: This table shows sample means with standard deviations in parentheses. School Pratibha Parv scores and denote means of the raw test scores at the school level in the respective years. Scores range from 0 to 100. Note that all schools available in the data for a given year are included here, while only the schools for which we have access to test scores in two adjacent years will be included in the RD analysis. The table is based on school level administrative dataset from 2012-2016.

Table 2: Reduction of discrepancy with improved exam procedures

	Math		Hindi	
Panel A	*Dep. var.: Difference in % correct (PP-Audit)*			
Treatment	-0.235***	-0.235***	-0.264***	-0.274***
	(0.047)	(0.047)	(0.025)	(0.025)
Constant	0.541***	0.541***	0.358***	0.360***
	(0.053)	(0.021)	(0.029)	(0.007)
School FE	No	Yes	No	Yes
Observations	1081	1080	1526	1525
Average discrepancy	0.44	0.44	0.29	0.29
Number of schools	49	48	49	48
Panel B	*Dep. var.: Correct response in independent test*			
Treat x Correct	0.090**	0.087**	0.075*	0.066
	(0.034)	(0.034)	(0.043)	(0.046)
Correct from PP	0.043*	0.042*	0.132***	0.144***
	(0.026)	(0.022)	(0.024)	(0.023)
Constant	0.213***	0.215***	0.380***	0.372***
	(0.008)	(0.009)	(0.022)	(0.017)
Item FE	Yes	Yes	Yes	Yes
School FE	No	Yes	No	Yes
Observations	4942	4942	5869	5869
Mean Audit % Correct	0.27	0.27	0.50	0.50
Mean official % Correct	0.69	0.69	0.80	0.80

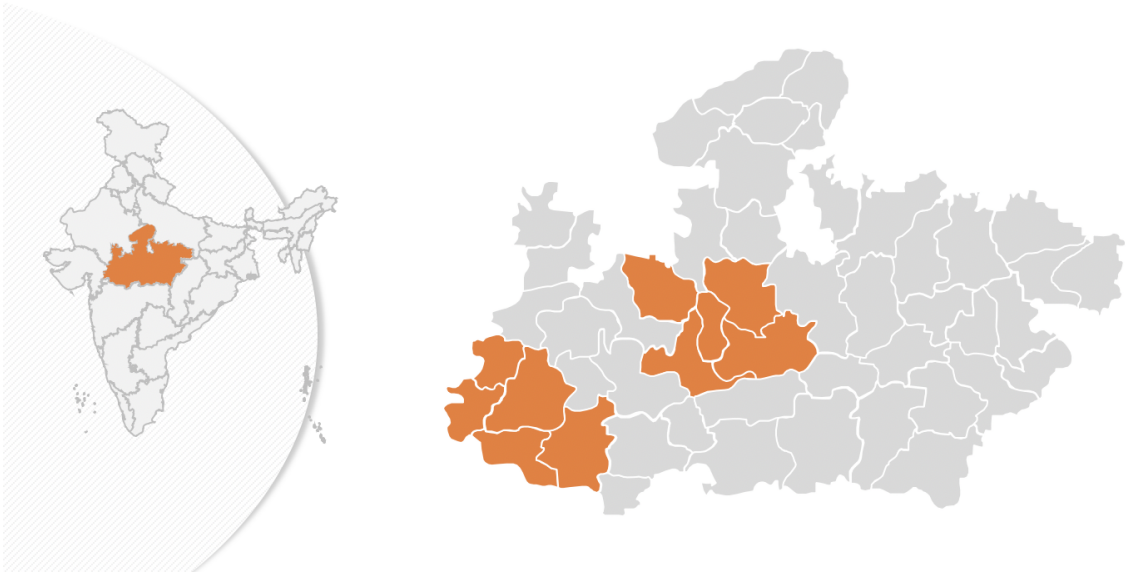
Notes: $p < 0.01$ = ***, $p < 0.05$ = **, $p < 0.1$ = *. Robust standard errors, clustered at school level, in parentheses.

Panel A: Each observation is at the student level. The dependent variable is the difference between the proportion of correct responses in the official and independent assessments to multiple-choice questions which were common across both tests. Treatment refers to Grade 8, which administered multiple test booklets and external grading to deter cheating. The sample is restricted to middle schools (Grades 6-8). In mathematics, there were no multiple-choice questions in Grade 6 that were common between the test and the retest and therefore the estimation sample only includes Grades 7 and 8.

Panel B: These regressions show the correspondence between responses to individual test questions across the official (Pratibha Parv) and independent tests by the same student. Each observation is at the item-level. The sample is restricted to multiple-choice questions administered in Grades 6-8 (middle schools) in both assessments. All specifications include item fixed effects: test questions are distinct across grades and therefore this also controls for grade level fixed effects. Treatment refers to Grade 8, which administered multiple test booklets and external grading to deter cheating. Being reported as answering an item correctly in the official test predicts answering correctly in the independent retest in all grades but differentially more so in the treated grade.

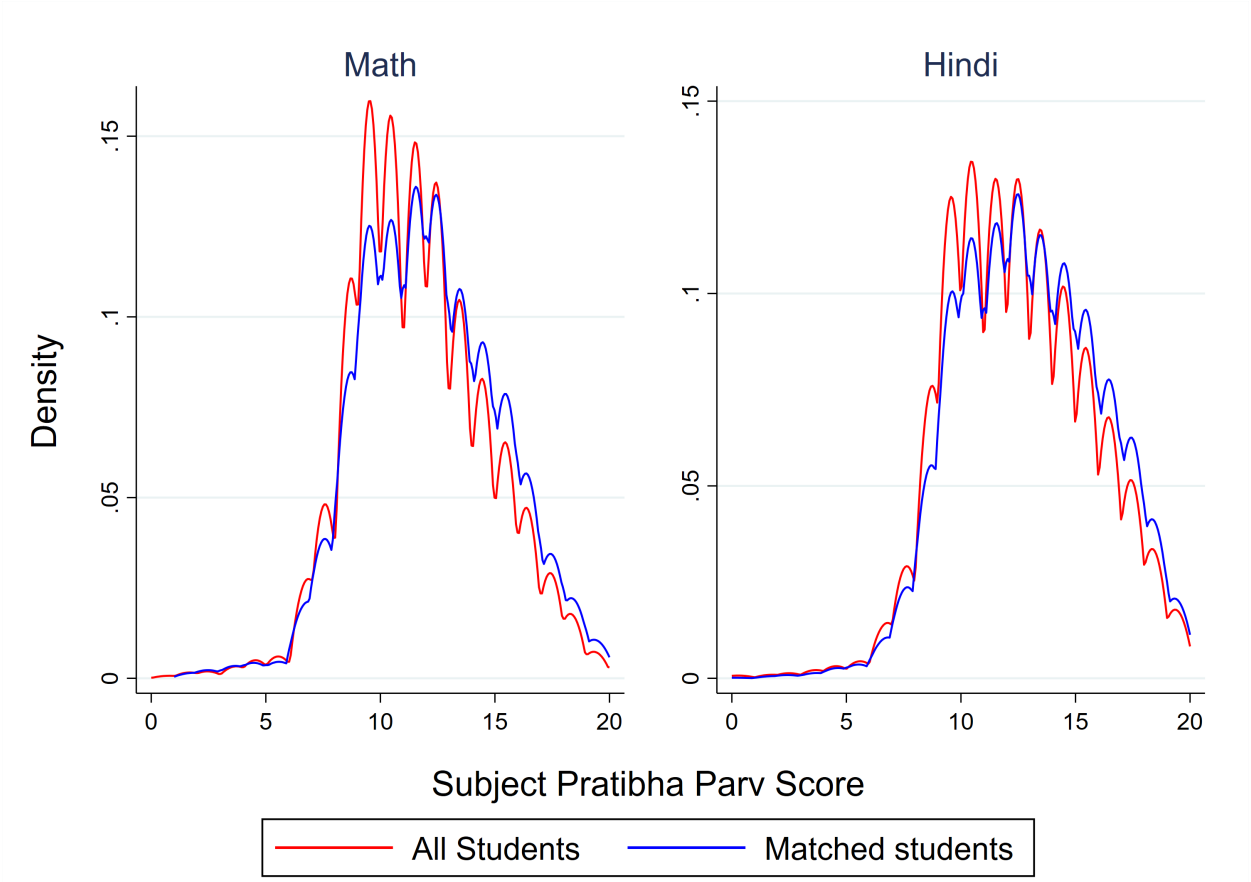
A Additional Figures and Tables

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



Notes: This graph compares the distribution of scores on the official test, standardized within grade and subject, for the full population of students to those for whom we also have independent assessment data (matched students). Matched students have modestly higher test scores but the distribution of achievement has substantial overlap and is very similar across the samples. The modest positive selection into retesting is expected since school attendance and achievement are positively correlated (and therefore, higher achieving students have a greater likelihood of being present on the day of the retest).

Table A.1: Sample characteristics of surveyed schools in Madhya Pradesh

	Whole state	Study districts	Sample
School level characteristics			
Enrolment in Elementary grades	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

Notes: The table presents mean values of observable school-level characteristics using year 2016-17 DISE Data and of school-level test scores using year 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 for which item-level data was collected. Standard deviations are reported in parentheses.

B Direct observations and qualitative interviews in M.P.

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 51 classrooms across 17 schools 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 (2020), 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 we restrict 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.