Out of Sight, Out of Mind? The Gap between Students' Test Performance and Teachers' Estimations in India and Bangladesh*

Sharnic Djaker† Alejandro J. Ganimian‡ Shwetlena Sabarwal§

Abstract

This is one of the first studies of the mismatch between students' test scores and teachers' estimations of those scores in low- and middle-income countries. Prior studies in high-income countries have found strong correlations between these metrics. We leverage data on actual and estimated scores in math and language from India and Bangladesh and find that teachers misestimate their students' scores and that their estimations reveal their misconceptions about students in most need of support and variability within their class. This pattern is partly explained by teachers' propensity to overestimate the scores of low-achieving students and to overweight the importance of intelligence. Teachers seem unaware of their errors, expressing confidence in estimations and surprise about their students' performance once revealed.

JEL codes: C93, I21, I22, I25

Keywords: teachers' estimations, differentiated instruction, India, Bangladesh

---

* We thank Noam Angrist and T. M. Asaduzzaman for inputs in the design and analysis of the Bangladesh results; Larry Aber, Minahil Asim, Felipe Barrera-Osorio, Rezarta Bilali, Andy de Barros, Emmerich Davies, Anne Fitzpatrick, Erin Godfrey, Diane Hughes, Isaac Mbiti, Abhiroop Mukhopadhyay, Mauricio Romero, Hiro Yoshikawa, and seminar participants for comments; and Rashmi Menon for excellent research assistance in the original India study. The study in India was funded by the Abdul Latif Jameel Poverty Action Lab's Post-Primary Education fund. The study in Bangladesh was funded by the Foreign, Commonwealth & Development Office of the Government of the United Kingdom, World Bank, and the China-World Bank Group Partnership Facility. The studies were reviewed and approved by the Institutional Review Boards at the Institute for Financial Management and Research (for India) and the Institute of Health Economics at the University of Dhaka (for Bangladesh).

† Doctoral Candidate, Psychology and Social Intervention, Steinhardt School of Culture, Education, and Human Development, New York University. E-mail: sharnic.djaker@nyu.edu.

‡ Assistant Professor of Applied Psychology and Economics, Steinhardt School of Culture, Education, and Human Development, New York University. E-mail: alejandro.ganimian@nyu.edu.

§ Senior Economist, The World Bank. E-mail: ssabarwal@worldbank.org.

Existing evidence from low- and middle-income countries (LMICs) suggests that teachers rarely cater to the needs of students at different achievement levels, and that this lack of differentiation disproportionately hurts low-achieving students, who are most out of step with curricular expectations, and thus with grade-based materials and instruction. Cross-sectional studies have found that many children cannot answer basic questions in arithmetic or reading by the end of primary education, suggesting that schooling is insufficient for them to acquire foundational skills (ASER, 2023; Banerjee et al., 2023; World Bank, 2018). Longitudinal studies have shown that learning improves far more slowly than expected, suggesting that early gaps in knowledge are seldom remedied as children move from primary to secondary education (Andrabi et al., 2007; Muralidharan et al., 2019; Singh, 2019). Classroom observations have revealed that teachers tend to use whole-classroom instructional approaches (instead of small-group or one-on-one instruction) and the same materials for all students (Bhattacharjea et al., 2011; Sankar & Linden, 2014).

The prevailing explanations for why teachers do not differentiate instruction focus on the incentives that they face. In many LMICs, national curricula were originally developed to prepare elites for post-secondary education and government jobs, and they are consequently overly ambitious, encouraging teachers to favor breadth over depth and leaving them little time to check for students' understanding (Pritchett & Beatty, 2015). Parents from low-income backgrounds often underestimate the returns to each additional year of schooling and pull their children out of school before they graduate, leading teachers to anticipate the pattern and underinvest in remediation (Banerjee & Duflo, 2011; Sabarwal et al., 2022). Bureaucrats, school principals, and parents place too much emphasis on students' performance on high-stakes exams during primary and secondary education, encouraging teachers to focus on maximizing pass

rates, and thus on the students who are most likely to take and pass these tests, at the expense of their peers (Duflo et al., 2011).

In this paper, we highlight another factor that may contribute to this pattern: most teachers in LMICs do not have an accurate understanding of the achievement of their students. We leverage two datasets on students' achievement (i.e., their scores on standardized tests of math and language) and teachers' estimates of their achievement (i.e., the scores teachers expected students to obtain) from primary and lower-secondary school (grades 5 and 6) in one city in India and three divisions in Bangladesh to document the extent of this misestimation and consider what factors may explain it.

We report six main sets of results. First, teachers misestimate their students' test scores. They tend to overestimate performance in math and underestimate it in language. If we calculate the difference between each student's test score and their teacher's estimation of that score, 84% of those differences in India and 60% of those in Bangladesh are positive in math (indicating teachers' estimations are above students' scores) and 66% of differences in Bangladesh are negative in language (reflecting estimations below scores). The magnitude of these errors is large in absolute terms (the mean difference between actual and estimated percentage-correct scores ranges from 7 to 24 percentage points [pp.], depending on the country and subject) and relative terms (the mean difference is between 151% and 218% of the within-class standard deviation [SD] in test scores).

Second, teachers' estimations reveal underlying misconceptions about which students are in greater need of support and how much students in the same classroom vary in their achievement. If we rank students within a class first according to their test scores and then according to their teachers' estimations of those scores, the correlation between these rankings

ranges from 0.07 to 0.35 across countries and subjects, implying that only 1 to 12% of variation in students' actual ranks is explained by teachers' estimated ranks. If we compare the within-class SDs in students' test scores to the within-class SDs implied by teachers' estimated scores, over half of teachers in India and three-fourths of those in Bangladesh underestimate variability in both subjects.

Third, teachers make different mistakes along the within-class achievement distribution. They are particularly likely to overestimate the performance of low achievers. In India, teachers overestimate the scores of students in all terciles in math, but they do so by a larger margin for the bottom tercile (30.3 pp.) than for the top tercile (15.9 pp.) In Bangladesh, teachers overestimate scores for the bottom tercile (by 18.6 pp. in math and 5.02 pp. in language) and they underestimate them for the top tercile (by 6.05 pp. in math and 17.9 pp. in language). In Bangladesh, we observe a similar pattern if we group students based on the national achievement distribution.

Fourth, the patterns above are largely explained by teachers overweighting the importance of students' intelligence for their achievement. In India, where we observe students' performance on an assessment of "fluid intelligence" (i.e., non-verbal abstract reasoning), the difference in math scores between students in the bottom and top terciles is much smaller than the one estimated by teachers. In fact, intelligence is a better predictor of teachers' estimated scores than of students' actual scores. Other student characteristics (e.g., being female, low-income, or from a scheduled caste or tribe) play a comparatively minor role in explaining teachers' misestimations. The metrics that school systems use to decide teacher pay (e.g., holding a master's degree, completing pre-service training, or having more years of experience) do little to explain variability in the accuracy of estimations.

Fifth, misestimations are not attributable to confounders. In Bangladesh, where we validated the patterns that we initially encountered in India, we are able to rule out several potential threats, such as teachers focusing on non-tested skills in their estimations (most teachers agree the tests measure what they teach), not recognizing their students (most teachers are able to verify the names of their students), not knowing how to estimate test scores (most teachers also underestimate the percentage of low-achieving students in their class), or not being able to aggregate judgments into total scores (teachers' estimations are no more precise for topics or specific items in the tests).

Lastly, teachers are largely unaware of their incorrect assessments of students' skills. The vast majority of teachers in Bangladesh express high levels of confidence in their estimations, and most admit being surprised when they learn how their students fare in the national distribution. The estimations of teachers with the highest level of confidence are most predictive of actual scores, but even among these teachers, estimations are quite inaccurate.

We make four main contributions. First, we advance global knowledge on the important question of whether teachers hold accurate beliefs about their students' skills. There is a large literature in education psychology on this question, but it is almost entirely based on studies in high-income countries (HICs),[1] where the correlation between students' test scores and teachers estimation is strong, ranging from 0.62 (Südkamp et al., 2012) to 0.65 (Hoge & Coladarci, 1989;

[1] According to a recent review of this literature, 70% of studies were conducted in the United States, 16% in Europe, 4% in Canada, 6% in Australia, and only 4% in other countries (Südkamp et al., 2012).

Kaufmann, 2020).[2] We are among the first to document that these correlations are much weaker in LMICs, examine the importance of locally relevant predictors (e.g., students' caste/tribe), and highlight differences in patterns across contexts (e.g., in HICs, students' fluid intelligence predicts their test scores more strongly than teachers' estimations; in LMICs, we find the opposite pattern).[3]

Second, we contribute to ongoing debates in development economics on the effectiveness of differentiated instruction (also known as "teaching at the right level"). Over the past decade, several randomized evaluations have found that interventions that encourage teachers to administer diagnostic tests, group students based on their performance, and assign different activities to each group improve test scores in LMICs (Banerjee et al., 2017; Banerjee et al., 2010; Banerjee et al., 2011; Banerjee et al., 2007; Duflo et al., 2020), but the relative contribution of each of their components remains unexplored. Our study offers suggestive

[2] Urhahne and Wijnia (2021) recently updated these reviews, but they did not report correlations. Instead, they reported a near-zero mean difference between students' scores and teachers' estimations, which is consistent with high correlations.

[3] Wadmare et al. (2022) compared students' achievement to teachers' estimations using data from India and found that 40% of teachers incorrectly perceived that their low-achieving students had achieved foundational literacy. Yet, their sample only included low achieving students, which precludes them from assessing whether estimations are *differentially* inaccurate for this group, and thus, to identify potential implications for how teachers allocate attention across groups. Further, their test only assessed foundational skills (e.g., in math, only number recognition, subtraction, and division), which understate the extent to which children lag behind curricular expectations (de Barros & Ganimian, 2023) and may accordingly underestimate the extent to which teachers are actually aware of gaps in students' knowledge and skills. Lastly, their test only categorized students into five performance levels (instead of producing a score), and thus can only indicate a relatively coarse level of correspondence between students' skills and teachers' estimations of those skills.

evidence that one mechanism through which these interventions work is by updating teachers' priors on the achievement of their students. We believe this hypothesis is consistent with the promising evidence on formative-assessment interventions in these contexts (de Hoyos et al., 2023; de Hoyos et al., 2021; de Hoyos et al., 2017).

Third, we add to growing evidence in international education on teacher capacity in LMICs. The earliest studies in this body of research documented low levels of teacher effort in these settings, as proxied by absences (Bruns et al., 2011; Chaudhury et al., 2006; Muralidharan et al., 2017). A subsequent wave of studies highlighted low levels of lesson time spent on instructional activities (Bruns & Luque, 2014; Stallings et al., 2014). Then, several studies drew attention to gaps in teachers' subject-matter and pedagogical expertise (Bold et al., 2017; Metzler & Woessmann, 2012; Santibañez, 2006). We build on recent studies on teachers' beliefs (Sabarwal et al., 2022), which find that many teachers do not believe it is their job to remedy learning from previous grades, to highlight the importance of teachers' understanding of students' skills for the frontier challenge of catering to increasingly large and heterogeneous student groups, the extent of which has been documented across several recent studies in LMICs (Ganimian & Djaker, 2022). Although our study is not designed to estimate the causal effect of teachers' estimations on their effort, and subsequently, on students' achievement, we identify this as a potential mechanism to be examined.

Finally, our results are in conversation with the literature on teachers' discrimination against students in LMICs. A number of studies have documented the prevalence of discrimination beliefs and behaviors using novel methods (Farfan Bertran et al., 2021; Hanna & Linden, 2012). We add some nuance to this discussion, comparing the crucial role that some attributes play in shaping teachers' estimations (e.g., students' fluid intelligence) to the relatively

less central influence of the factors that have been previously highlighted (e.g., students' sex, income, and caste/tribe).

The rest of the paper is structured as follows. Section 2 offers an overview of the methods of the studies in India and Bangladesh and highlights several aspects of their design that draw on lessons from prior research to maximize the accuracy of teachers' estimations. Section 3 presents the main results of these studies. Section 4 concludes by drawing general lessons from our findings, situating our studies in the broader literature, and discussing promising directions for future research.

2. **Methods**

In this paper, we draw on data from two studies in India and Bangladesh. The study in India, in which we reanalyzed data from a randomized evaluation, first allowed us to identify the mismatch between students' test scores and teachers' expectations. The one in Bangladesh, which we designed and executed, then enabled us to probe this pattern, rule out confounders, and identify potential mechanisms. We present them together because, given the similar incentive structures and capacities across these two school systems, we see their consistent results as suggestive of common challenges.

Sample

In study 1, we use data from a randomized evaluation of a teacher-residency program in the state of Maharashtra, India (Ganimian et al., 2024). This evaluation was conducted in grades 5 and 6, in a convenience sample of 48 English-medium public primary schools in the city of Pune, in the Indian state of Maharashtra. The sampling frame included all 286 primary schools

run by the Pune School Board.[4] We only use data from the control group (to eliminate the possibility that the intervention may influence the patterns we observe) for regular teachers for the endline round of data collection.[5] Our data-analytic sample includes 46 teachers and 457 students from grade 5 and 6.

In study 2, we use data from a study that we conducted in Bangladesh to understand the extent to which the patterns we observed in India were present in a school system with similar institutional features. This study was conducted in grade 6, in a stratified random sample of 403 math and/or language (Bangla) teachers and their 1,306 students across 273 secondary public-private partnership (i.e., publicly funded, privately managed or PPP) schools, which account for more than 95% of enrollment in secondary education. We arrived at this sample as follows. First, we obtained access to the results of the 2019 National Assessment of Secondary Students (NASS)—a low-stakes large-scale assessment of a nationally representative sample of students in math, Bangla, and English. This dataset included 28,238 students, whom we matched to their 6,373 teachers across all three tested subjects. Then, we drew a simple random sample of three of the eight divisions in the country (Chittagong, Dhaka, and Mymensingh) and we kept the 2,724

[4] The authors excluded 118 schools in remote rural areas (because it would have been challenging to monitor the intervention in those contexts), 30 Urdu-medium schools (because most of the teaching residents did not speak Urdu), 46 English-medium schools and 13 schools that were implementing other interventions (to avoid confounding the effects of the intervention with these other programs), 20 schools with low enrollment (to minimize sampling error), and 9 schools that had already implemented the intervention. After baseline, two schools were dropped due to a shortage of residents.

[5] The treatment group was taught by college students during a one-year fellowship (instead of by public-school teachers). We are interested in explaining the behavior of public-school teachers, which is why we excluded the treatment group.

teachers in those divisions. Lastly, we excluded 1,034 teachers who did not teach math or Bangla, 83 teachers for whom we did not have cell phone numbers, and 84 teachers for whom we did not have class-size information (which we needed for stratification). From the remaining 1,523 teachers, we randomly sampled 825, stratifying our selection by terciles of student enrollment. Of the sampled teachers, we included 573 in our sample and allocated the remaining 252 to a back-up roster. We called 726 teachers (573 from the sample and 153 from the back-up roster), reached 607 of them, obtained consent from 597, excluded 194 (because they reported that they did not teach in the target schools, grades, or subjects), and were left with 403. We offered each teacher BDT 100 taka (~USD 1.17) in cell-phone credit to participate. We matched these teachers to 3,259 students, of whom they recognized 2,445. We randomly selected 1,128 of these students to elicit teachers' estimations of their test scores.

Data

In study 1, we mainly use data from two instruments administered at the endline of the impact evaluation: a math test and a survey of teachers. The math test used 35 multiple-choice items to assess students on three content domains (numbers, geometry and measures, and data display) and cognitive domains (knowing, applying, and reasoning) across a wide range of difficulty levels.[6] Each student's score is given by the percentage of items that they answered correctly, from 0 to 100. The survey asked teachers to estimate the scores of 10 randomly

[6] The distribution of items across content and cognitive domains was based on the assessment framework of the 2019 Trends in International Math and Science Study (TIMSS; IEA, 2017). The items for the test were drawn from domestic and international assessments, as well as from impact evaluations conducted in India. The impact evaluation for which this test was designed found moderate-to-large effects on this test at endline and follow-up (Ganimian et al., 2024).

selected students in their classroom. Each estimation is also expressed as a percentage-correct score from 0 to 100. We also leveraged data from the baseline to explore whether the accuracy of estimations varied by students' characteristics (being female, from a low-income family, or from a scheduled caste or tribe) or teachers' characteristics (e.g., having a master's degree, pre-service training, or more experience).

In study 2, we mainly use data from two sources: the 2019 NASS for grade 6, which assessed students' skills in math and language (Bangla), and a phone-based survey of teachers that elicited their estimations of their students' NASS performance. The 2019 NASS used 40 items (of which 29 were multiple choice) across five content domains in math (number and operations, measurement, data, geometry, and algebra) and 42 items (of which 36 were multiple choice) across three content domains in language (reading, vocabulary, and grammar). The distribution of items was based on the national curriculum. The survey asked teachers to estimate the scores of three randomly selected students in their classroom on the test assessing the subject that they taught (math or language).[7] It also included questions on teachers' characteristics (e.g., education), beliefs (e.g., growth mindset), and instructional practices, which we used to explore heterogeneity in the accuracy of estimations.

There are three aspects common to both studies that draw on lessons from prior research and maximize our chances of finding a relationship between students' test scores and teachers' estimations. First, we asked teachers to estimate their students' scores on a *specific test*, instead of asking them to provide a holistic appraisal, to avoid scores and estimations differing simply because the latter drew on information not captured by the test. Second, we asked teachers to

[7] 212 teachers taught math, 181 taught Bangla, and 10 taught both and were randomly assigned to one for the survey.

express estimations using the *same scale* on which tests were scored, instead of asking them to use a Likert scale, to prevent mismatches between scores and estimations purely due to differences in scales. Third, we asked teachers to estimate each *individual student's score*, instead of ranking their students, to rule out the possibility that differences between scores and estimations were attributable to misjudgments about students' relative standing (for reviews discussing the importance of these factors, see Hoge & Coladarci, 1989; Südkamp et al., 2012; Urhahne & Wijnia, 2021). Lastly, we asked teachers to provide estimations after they had taught students for a full school year, to avoid the possibility that estimations may be inaccurate because teachers had not gotten to know students.

Each study also has some specific features that allow us to rule out potential confounders in the relationship between students' scores and teachers' estimations. In India, we measured students' achievement and teachers' estimations at the *same time* (i.e., within the same round of data collection, often on the same day) to avoid discrepancies due to "recall bias" (i.e., teachers misremembering their students' test performance) or "mean reversion" (i.e., teachers anticipating changes in their students' scores over time).[8] Given that the test was developed by a research team for the purposes of an impact evaluation, we also showed teachers the test we used to assess students before we elicited their estimations to keep them from making mistakes simply because they were unfamiliar with the test.[9] In Bangladesh, we measured students' test scores and

[8] In Bangladesh, student achievement was measured by the 2019 NASS in February-March of 2020 and teachers' estimations were elicited by our survey in September of 2020, so five months elapsed between both measurements.

[9] In Bangladesh, we asked teachers to estimate their students' scores on the national assessment, which has been administered since 2011 and with which teachers were already familiar. The test assesses the material in the grade 6 curriculum and is thus ought to be aligned with the material teachers cover in class (MEW-DSHE, 2019).

teachers' estimations in two subjects to rule out discrepancies attributable to estimating performance in a specific subject.[10] We also asked teachers to estimate students' performance not just on the test as a whole, but also on specific topics and items to prevent the possibility that mismatches originated from teachers not knowing how to estimate overall test performance.[11] Lastly, we asked teachers to corroborate the names of their students, using both real and fake names, to rule out the possibility that they did not know the students for whom they were asked to estimate scores.[12]

3. Results

In this section, we present the findings from the studies in India and Bangladesh. We document the average gap between students' test scores and teachers' estimations, explore how this gap differs across sub-groups of students and teachers, and rule out potential confounders.

A. Average Gaps between Students' Test Scores and Teachers' Estimations

Teachers Misestimate the Test Scores of Their Students

All teachers in our data misestimate the performance of their students on standardized tests. If we calculate the difference between each student's percentage-correct score and their teacher's estimation of that score, only 2 (of 438 teacher-by-student) estimations are exactly 0 (indicating that the teacher guessed a student's score perfectly) in India for math, 22 (of 605 estimations) in Bangladesh for math, and 2 (of 509 estimations) for language. None of the teachers in either country estimated the scores of all of their students perfectly (as explained in

[10] In India, the dataset only includes teachers' estimations of students' test performance in math.

[11] In India, the data only includes teachers' estimations of students' overall scores.

[12] In India, both students and teachers had been part of a year-long study, so such verification was not necessary.

the Methods section, in India, we asked teachers to estimate the scores of 10 students, and in Bangladesh, of three students, so this fact is less surprising in India than in Bangladesh).

The direction (i.e., sign) of teachers' errors varies by subject. In math, teachers tend to overestimate students' test performance. If we tally the differences between students' scores and teachers' estimations in this subject, 84% of those differences are positive (indicating that the teacher guessed a score that was higher than the student's actual score) in India and 60% are positive in Bangladesh. In language, the opposite is true: only 33% of differences in that subject are positive in Bangladesh. We show this pattern graphically in Figure 1, which displays the distribution of differences between scores and estimations at the student-teacher (dyad) level (in histograms) and the distribution of the teacher-level averages of those differences (in density plots) by country and subject. The distributions are centered to the right of 0 in math and to the left of 0 in language, which is consistent with the direction of teachers' errors above (see also Figure A.1 in the appendix).[13]

The Magnitude of Teachers' Errors is Large

The simplest way to quantify the discrepancies between students' test scores and teachers' estimations is to calculate the average difference in percentage points between these measures, given that both are expressed in percent-correct terms (i.e., from 0 to 100). In India, the mean actual score in math was 39, but the mean estimated score is 63—24 pp. higher. In Bangladesh, the mean actual scores are 56 in math and 70 in language, but the mean estimated

[13] The differences in the direction and magnitude of misestimations across subjects could be due to a number of factors, including differences in the difficulty of estimating performance in each subject, in the difficulty of the tests, or in the alignment between the material taught in class and the one assessed by the test.

scores are 65 and 63—8.4 pp. higher in the first case and 7 pp. lower in the second case (Table A.1 in the appendix).

One way to make sense of these differences is to calculate the percentile rank of the average estimated score in the distribution of actual scores. A difference of 0 would indicate that the mean estimated score is at the 50th percentile of distribution of actual scores. The average difference in India indicates that the mean estimated score is at the 88th percentile of the actual-score distribution; the average differences in Bangladesh indicate that the mean estimated scores are at the 58th and 57th percentiles of the actual-score distributions for math and language, respectively.

The approaches above, however, do not take into account that some teachers have students who vary more in their test performance than others. This is an important shortcoming because if two teachers misestimate the scores of their students by the same amount in percentage points, but one of them has students who vary more in their scores, that teacher will be less able to distinguish between the scores of any two students. An alternative approach that illustrates the implications of misestimations for instruction is to divide each teacher's average difference between the estimated and actual scores of students in their class by that teacher's within-class SD in actual scores and to calculate the mean ratio across all teachers. If we do so, the average difference between estimated and actual scores is equivalent to 1.51 within-class SDs in India for math, and to 2.18 and 1.87 within-class SDs in Bangladesh for math and language, respectively (Table A.1).

The most frequently used metric in prior studies (which, as explained in the Introduction, were conducted primarily in high-income countries) is the correlation between actual and estimated scores. The Pearson correlation coefficient between these scores is 0.36 in India for

math (which is equivalent to an R-squared of 0.13, implying that only 13% of variation in actual scores is explained by variation in estimated scores), 0.07 and 0.11 in Bangladesh for math and language, respectively (equivalent to R-squareds of approximately 0.01 or 1% of variation in actual scores explained by variation in estimated scores). We show these relationships graphically in Figure 2. In all cases, the line of best fit is flat, indicating a positive but weak association between actual and estimated scores.

A similar approach is to estimate the extent to which estimated scores predict actual scores. If we fit an ordinary least-squares (OLS) regression of actual on estimated scores, we can obtain the average change in estimated scores for every one-point change in actual scores. In India, a one-point increase in actual scores corresponds to a 0.31-point increase in estimated scores for math ($p<0.01$); in Bangladesh, it corresponds to 0.10- ($p=0.09$) and 0.13-point ($p=0.02$) increases in estimated scores for math and language, respectively (Table A.2, column 1). If we include teacher fixed effects (to account for differences in the distributions of actual and estimated scores across teachers), the point estimates increase to 0.45 ($p<0.01$) in India for math, and 0.32 ($p<0.01$) and 0.44 ($p<0.01$) in Bangladesh for math and language, respectively (Table A.2, column 2). Fixed effects also increase the share of variation in actual scores due to variation in estimated scores, partly due to non-random assignment of students to teachers (e.g., some teachers are assigned to lower-achieving students).

Teachers Do Not Know They Are Making Mistakes

Teachers are not simply unaware that they are making incorrect estimations of students' test scores; they are actually overwhelmingly confident in their erroneous judgments. In Bangladesh, we asked them to indicate their level of confidence in their estimations of individual students' scores, using a four-point scale ranging from "Certain (I expect all my scores to be

correct within a few percentage points)" to "Not confident (I am mostly guessing based on my experience)". The vast majority of teachers of both math (91%) and language (83%) selected the top two levels of the scale, indicating either that they were certain or very confident in their estimations. In fact, not one of the 403 teachers surveyed in either subject indicated that they were not confident in their estimations.

The most confident teachers estimate students' scores more accurately than the rest, but only by a small margin, and other levels of confidence are less helpful in predicting estimation accuracy. If we regress actual on estimated scores separately for each level of confidence selected by at least one teacher in Bangladesh (in descending order: certain, very confident, and somewhat confident), the estimations of teachers with the highest level of confidence are most predictive of actual scores. Yet, even among these teachers, estimations are quite inaccurate: in math, a one-point increase in actual scores corresponds to a 0.32-point increase in estimated scores ($p<0.01$); in language, it corresponds to a 0.25-point increase ($p<0.01$; Table A.3). Alternatively, only about 5% of variation in actual scores is explained by variation in estimated scores in both math and language. Lower levels of confidence correspond to lower levels of accuracy in math (among teachers who are very confident, a one-point increase in actual scores corresponds to a 0.07-point increase in estimated scores [$p=0.346$], and among those who are somewhat confident, to a 0.12-point *decrease* [$p=0.544$], suggesting that estimated scores are inversely related to actual scores), but less clearly in language (the corresponding figures are 0.04- [p=0.687] and 0.14-point [$p=0.274$] increases, respectively)—partly, because a very small share of teachers selected the lowest level of confidence.

Teachers Do Not Know How Their Students Compare

Teachers may not need to know the exact score of each of their students to understand that some are struggling more with the material, and are thus in greater need of support, than others. Yet, they seem to be unaware of the within-class relative standing (i.e., ranking) of students in their classrooms. The simplest way to show this is to calculate the Spearman correlation (i.e., correlation in ranks) between actual and estimated scores.[14] These correlations are very similar to the Pearson correlations above: 0.35 in India for math (which is equivalent to an R-squared of 0.12, implying that only 12% of variation in actual ranks is explained by variation in estimated ranks), and 0.07 and 0.11 in Bangladesh for math and language, respectively (equivalent to R-squareds of 0.01 and 0.02 or 1% and 1.5% of variation in actual ranks explained by variation in estimated ranks, respectively). We plot the relationship between students' actual within-class ranks (as indicated by the relative standing of their score in the class distribution) and teachers' estimated within-class ranks (as indicated by the relative standing implied by each estimated score) in Figure 3. Again, the lines of best fit are flat, indicating positive but weak associations.

We confirm this pattern with a regression framework. If we fit an OLS regression of actual on estimated within-class ranks (as defined above), we can obtain the average change in estimated ranks for every one-point change in actual ranks. In India, a one-point increase in actual ranks corresponds to a 0.51-point increase in estimated ranks for math ($p<0.01$); in

[14] We compare the rankings of students' scores and teachers' estimations for students who have both. As explained in the Methods section, in India, we asked teachers to estimate the scores of 10 students, and in Bangladesh, of three students. We do not expect the number of students selected in each country to matter for rankings because we did not ask teachers to rank only selected students. Rather, we are ranking those students according to their actual and estimated scores.

Bangladesh, it corresponds to a 0.17-point ($p<0.01$) *decrease* (indicating an inverse relationship between actual and estimated ranks) for math and a 0.14-point ($p=0.02$) increase in language (Table A.4). The R-squareds are higher than those for the regressions of scores (0.25 in India for math and 0.03 and 0.02 in Bangladesh for math and language, implying that 6% and less than 1% of variation in actual ranks is attributable to variation in estimated ranks, respectively) due to the coarse nature of rankings (even if a teacher misestimates a student's score, they may still estimate that student's rank correctly).[15]

In fact, a non-trivial share of teachers *flip* their students' rankings (i.e., they guess the top-scoring student is the bottom-scoring student or vice versa). In Bangladesh, where we only elicited teachers' scores for three students, we calculate the percentage of teachers who fall in this category: 11% of teachers in math and 12% in language reversed the relative standing of their own students.

Teachers Underestimate How Much Their Students Vary

Even if teachers do not know the absolute or relative test performance of their students, they may still have a sense of how much their students vary in their understanding of the material.[16] Yet, if we compare actual within-class variability (the within-class SD in students' test scores) to the estimated within-class variability (the within-class SD in teachers'

[15] Including teacher fixed effects makes almost no difference in our results, given that the distribution of rankings does not vary within country (recall that we elicited estimations for 10 students in India and three students in Bangladesh).

[16] For example, if in a two-student classroom one student scores a 0 on a test and the other scores a 100, and their teacher estimates the first student to score a 100 and the second student to score a 0, that teacher would be incorrect in their estimations of their students' scores, but they would be correct in their estimation of test-score variability in their class.

estimations), most teachers underestimate variability in their classrooms. In India, 54% of teachers underestimate within-class variability for math; in Bangladesh, 77% of teachers do so for math and 78% do so for language.

Teachers' beliefs about test-score variability in their classroom are a poor predictor of actual within-class variability in students' scores. The Pearson correlation coefficient between actual and estimated within-class variability (as defined above) is 0.08 in India for math (implying that about 1% of actual variability is explained by estimated variability), and 0.08 and -0.03 in Bangladesh for math and language (implying that less than 1% of actual variability is explained by estimated variability). We display these relationships graphically in Figure 4. In all cases, the line of best fit is flat, indicating a near-zero association between actual and estimated within-class variability. We confirm this pattern analytically: when we fit an OLS regression of actual on estimated within-class SDs, the coefficients are consistently around zero and statistically insignificant (Table A.5).

B. Gaps by Student and Teacher Characteristics

Teachers Overestimate the Scores of Low Achievers

The pattern presented above is partly explained by teachers being particularly likely to overestimate the scores of low-achieving students. We classify students into low-, middle-, and high-scoring groups in two ways. First, we group students according to their *within-class* tercile (i.e., relative standing in their class).[17] In India, teachers overestimate the scores of students for

[17] In India, we assign students to within-class terciles based on the scores of the 10 students for whom teachers provided estimations. In Bangladesh, we do so based on *all* students in the same classroom who took the national assessment, but we only observe both actual and estimated test scores for the three students for whom teachers provided estimations.

all terciles in math, but they do so by a larger margin for students in the lowest tercile (30.3 pp., $p<0.01$) than for those in the middle and highest terciles (23.6 and 15.9 pp., respectively, $p<0.01$; Figure 5 and Table A.6). In Bangladesh, teachers overestimate the scores of students in the lowest tercile (by 18.5 pp. in math and 4.84 pp. in language, $p<0.01$) and they underestimate the scores of those in the highest tercile (by 6.17 pp. in math and 17.9 pp. in language, $p<0.01$).

In Bangladesh, we also group students according to their *national* tercile (i.e., standing among all test takers). When we do so, we find the same pattern as above: teachers overestimate the scores of students in the lowest tercile (by 30.4 pp. in math and 14.0 pp. in language, $p<0.01$) and they underestimate those of students in the highest tercile (by 13.0 pp. in math and 20.1 pp. in language, $p<0.01$; Table A.7), suggesting that teachers underestimate low performance in their class.

Teachers Overweight the Importance of Intelligence for Test Scores

The gap between test scores and estimations is also partly explained by teachers overweighting the importance of students' intelligence when estimating their achievement. In India, we have data on students' "fluid intelligence" (i.e., non-verbal abstract reasoning) from an abridged version of Raven's Progressive Matrices (Raven & Summers, 1986). If we calculate the average math test score by quartile of students' intelligence, those in the quartile 1 (lowest) score lower (30.7) than those in quartile 4 (highest; 44.8)—a difference of 14.1 pp. Yet, teachers estimate the difference between these quartiles to be much larger—about 23.4 pp.—because, even if they overestimate students' test scores across the intelligence distribution, they overestimate the test scores of quartile 4 students by a larger margin (30.3 pp., $p<0.01$) than those of quartile 1 students (21.0, Table A.8). In fact, students' intelligence is a better predictor of teachers' estimations than of students' test scores: the Pearson correlation coefficient of

intelligence and test scores is 0.29 (implying an R-squared of 0.09) and the coefficient of intelligence and estimations is 0.36 (implying an R-squared of 0.13). These results suggest teachers use intelligence as a proxy for achievement.

Teachers Overestimate Girls' Test Scores by a Larger Amount Than Boys' Test Scores

Several studies in LMICs have found that teachers often hold different expectations for different groups of students (e.g., Farfan Bertran et al., 2021; Hanna & Linden, 2012). Therefore, we investigated whether teachers' estimations differed in levels or accuracy for traditionally disadvantaged students. We can only do so in India, where we observe students' demographics.

In India, male and female students perform similarly in math (with average percentage-correct scores of 39.4 and 38.7, respectively; Table A.9). Teachers do not estimate the scores of female students to be lower than those of males. In fact, they overestimate the scores of both male and female students, but they do so by a wider margin for female students (26.8 pp., $p<0.01$) than for male students (22.1 pp., $p<0.01$). Male teachers overestimate the scores of both male (28.5, $p<0.01$) and female (30.6, $p=0.01$) students by a wider margin than female teachers (20.7 and 25.9 pp., respectively, $p<0.01$), but the difference between estimations for male and female students is larger for female (5.2 pp., favoring females) than for male teachers (2.1 pp., also favoring females).

Teachers Overestimate the Scores of Low- and High-income Students by Similar Amounts

In India, students perform similarly in math across the socio-economic spectrum. If we group students according to their quartile in an index of household assets,[18] students in quartile 1 (poorest) perform comparably (with an average percent-correct score of 38.6) to those in quartile

[18] The index indicates whether students reported having the following assets at their homes: a desk to study, storybooks, a room of their own, a dictionary, a computer, a television set, Internet connection, and a DVD player.

2 (39.3), quartile 3 (39.3), and quartile 4 (richest, 39.2). Teachers do not estimate the scores of students from low-income families to be lower than those of high-income families. In fact, they overestimate the scores of students across all quartiles by a similar margin (between 22.7 and 28.3 pp., Table A.10).

Teachers Overestimate the Scores of Students of Different Castes/Tribes by Similar Amounts

In India, students perform similarly in math across castes and tribes. If we categorize students according to the caste or tribe registered in their school's records, those from scheduled tribes perform best (with an average percentage-correct score of 48), followed by those of other backward castes (41.9), other general categories (39.6), and those from scheduled castes (36.7).[19] The rankings of teachers' estimations mirror those of students' scores: they are highest for student from scheduled tribes (with an average estimated score of 71.6), followed by those of other backward castes (70.0), other general categories (64.1), and scheduled castes (61.9). Yet, estimations surpass scores for all groups by between 23.6 and 28.1 pp. ($p<0.01$ except for scheduled tribes; Table A.11).

Teachers with More Education, Training, and Experienced Are Not More Accurate

Given the weak association between the metrics that school systems in LMICs frequently used to determine teachers' salaries—holding a master's degree, completing pre-service training, or having more years of teaching experience than the median teacher—and student achievement gains (Araujo et al., 2016; Azam & Kingdon, 2015; Bau & Das, 2017; Buhl-Wiggers et al., 2019), it is perhaps unsurprising that these factors do not account for differences in the accuracy of estimations. If we fit an OLS regression of actual on estimated scores, an indicator variable for a characteristic (e.g., having a master's degree), and its interaction with estimated scores, we find

[19] The "general" category includes all non-disadvantaged classes as per the Indian caste system.

that more educated, trained, and experienced teachers do not make more accurate predictions.[20] In fact, the coefficients on the interactions in these regressions (which capture the differences in predictive power of estimations between teachers with and without a characteristic) are consistently estimated to be at zero (indicating no difference in predictive power with and without a characteristic) or below (indicating that teachers without the characteristic estimate scores *more* accurately; Table A.12).

C. Ruling Out Potential Confounders

Errors Are Not Due to Focus on Non-tested Skills

Teachers' errors cannot be due to them basing their estimations on information about students' skills on a subject that are not measured by the tests. We did not ask teachers to offer a holistic assessment of their students' skills; we asked them to estimate their test scores. In fact, most teachers believe that the tests measure precisely what they are teaching. In Bangladesh, we asked teachers "to what extent do you think [the test] covers what you are teaching?"[21] In math, 96% of them claimed that it matched what they taught to a "high" or "moderate" extent; just 4.2% selected a "low" level of alignment. In language, the corresponding figures were 96% and 3.8%, respectively.

Errors Are Not Due to Teachers Not Recognizing Their Students

Teachers' mistakes are not due to them not knowing or misremembering their students. In Bangladesh, we presented each teacher with the names of 12 students and asked them to verify whether they taught each of them. Two of those names were fake to check whether teachers

[20] We also analyzed the importance of teachers' experience at their school, at their grade, and at their subject. All of these results were consistent with those for overall experience, so we only kept those ones for brevity.

[21] The study in India did not include this question.

claimed to recognize students who were not in their classroom. Teachers were far more likely to recognize the real names (the average teacher recognized 75% of them) than the fake names (the average teacher recognized only 30% of these names and just 14% of teachers indicated knowing *both* fake names; Table A.13). Further, teachers were just as likely to recognize students in the top, medium, and bottom test-score terciles of the within-class and national achievement distributions.

Errors Are Not Due to the Challenge of Estimating Test Scores

Teachers' underappreciation of the extent of underperformance in their classroom cannot be explained by the difficulty of estimating test scores. In Bangladesh, we also asked them to estimate the percentage of their students in the top, middle, and bottom terciles of the national achievement distribution. Given that we observe the scores of all students who took the national student assessment, we can compare teachers' estimations to the actual percentages. Consistent with teachers' propensity to overestimate the test scores of low achievers (see Figure 5 and Table A.7), teachers are more likely to underestimate the percentage of these students in their classroom. The average teacher underestimates the share of students in the lowest tercile by 19.2 pp. ($p<0.01$) in math and by 14.5 pp. ($p<0.01$) in language (Table A.14). Teachers also underestimate the share of students in the middle tercile, but by a smaller margin: 4.26 pp. ($p=0.05$) in math and 3.53 pp. ($p=0.13$) in language. And while teachers underestimate the share of students in the top tercile in math by 3.86 pp. ($p=0.17$), they overestimate it in language by 5.43 pp. ($p=0.08$).

In fact, most teachers admitted to being surprised once they learned of the actual percentage of students in the lowest, middle, and top terciles of the national distribution in their classroom. When we revealed this information to them, 63% of math teachers and 60% of

language teachers indicated that they were "very" or "somewhat surprised." If there is a reluctance by survey respondents to admitting to making mistakes, these percentages may understate teachers' surprise.

Errors Are Not Due to the Difficulty of Aggregating Item Performance

Teachers' mistakes cannot be attributed to the level of aggregation of their estimations. To explore whether teachers made mistakes due to the difficulty of estimating total scores, in Bangladesh, we also asked teachers to estimate the percentage of students in their classroom who are proficient in the tested topics in math (number and operations, measurement, data, geometry, algebra) and language (reading, vocabulary, and grammar).[22] Consistent with our results for total scores (see Figure 1 and Tables A.1-A.2), teachers overestimate proficiency rates for all topics (Figure 6). The differences between estimated and actual percentages of proficient students by topic are large. In math, they range from 43.7 pp. for measurement to 56.1 for algebra (both $p<0.01$). In language, they range from 19.5 pp. for grammar to 57.7 pp. for reading (both $p<0.01$), indicating that the level of aggregation of the estimations does not explain their inaccuracy (Table A.15).

Teachers' estimations are no more accurate when they focus on specific items. In Bangladesh, we asked teachers to estimate the percentage of students in their classroom who would could answer *specific items* from the national student assessment (e.g., in math, calculating the area of a rectangle). Teachers do not overestimate performance on all math items and underestimate performance on all language items, as we might expect given the results above. Instead, the direction (i.e., sign) of their errors is related to the difficulty of the items. We

[22] We asked teachers to categorize a student as proficient on a topic if they were "able to answer *all* items on that topic". The study in India did not include this question.

identify this pattern by leveraging the dataset of the national student assessment, which includes item-level performance not just the students in our study, but for *all* test takers, and use a two-parameter logistic item-response theory (2PL IRT) model to estimate each item's difficulty (*b*) parameter (i.e., the level of latent ability required to answer each item with a 50% chance; see Yen & Fitzpatrick, 2006). Across both subjects, teachers tend to underestimate very easy items (i.e., those with low *b*-parameters in the IRT models), overestimate very hard items (i.e., those with high *b*-parameters; Figure 7). In fact, the magnitude of teachers' errors correlates with the magnitude of items' difficulty (i.e., errors are largest for the easiest and hardest items and smallest for those of average difficulty; Table A.16).

Teachers' estimations are no more accurate when they focus on specific items *and* students. In Bangladesh, we asked teachers to indicate whether three students whom they had previously recognized could correctly answer specific items in the math and language tests. Consistent with our results above on teachers' propensity to overestimate the test scores of low-achieving students (see Figure 5 and Tables A.6-A.7), underestimate the percentage of such students in their classroom (see Table A.14), and overestimate performance on easier items (see Table A.16), teachers overestimate the performance of low achievers and underestimate the performance of high achievers in both items presented for each math and language, regardless of whether students are classified into terciles based on the within-class or national achievement distributions (Table A.17).[23]

Errors Are Not Due to Overreliance of Heuristics

[23] Given that most items presented to teachers in this question were easy, we cannot answer the important question of whether teachers are more likely to overestimate the performance of low-achieving students on easier items. The results for the vocabulary item (*b=2.28*), however, suggests otherwise.

Teachers—like everyone else—rely on heuristics when making estimations. Many teachers estimate students' percentage-correct scores to be at or around 60%, a widely used passing rate: 11% of math estimations in India and 21% and 19% of math and language estimations in Bangladesh, respectively, are exactly at 60%; and 17% of those in India and 29% and 32% of those in Bangladesh are within 5 pp. of 60% (see Figure 2). Yet, as Table A.18 shows, even if we omit teachers with estimations at or around 60%, we observe the same pattern as above (see Table A.2).

4. Conclusion

In LMICs, teachers rarely differentiate their instruction. Prior studies have highlighted the *incentives* that teachers face, which discourage them from supporting these students. The present study complements such research by highlighting the potential role of *capacity*—specifically, the mismatch between students' test scores and teachers' estimations of those scores.

The study distinguishes itself from similar efforts by documenting common patterns across two contexts (India and Bangladesh), subjects (math and language), types of tests (one developed by an independent research team and one administered as part of a national assessment); examining heterogeneity in the accuracy of teachers' estimations across student and teacher characteristics; and ruling out potential confounders (e.g., eliciting teachers' estimations at the subject and item levels; see Tables A.19 and A.18 for an overview of main results and robustness checks, respectively).

We draw three main lessons from this research. First, teachers in LMICs differ considerably from their counterparts in HICs in their capacity to estimate their students' achievement. The correlations between actual and estimated scores that we document in India

and Bangladesh are two and nine times weaker than those in HICs, respectively (Hoge & Coladarci, 1989; Kaufmann, 2020; Südkamp et al., 2012; Urhahne & Wijnia, 2021). Second, teachers in LMICs are particularly prone to overestimate performance in the lower end of the achievement distribution. This pattern emerges most clearly in their overestimates of the scores of students in the lowest within-class and national achievement terciles, which has also been found in HICs (Begeny et al., 2008; Begeny et al., 2011), but it also manifests itself in teachers' proclivity to overestimate proficiency rates on easier topics and items. Lastly, teachers' misestimations seem to be more closely related to students' intelligence than with their attitudes towards student sub-groups (e.g., female or low-income). Studies in HICs do not typically find a relationship between teachers' estimations and students' intelligence (Hoge & Butcher, 1984), but they have found that estimations are rarely predicted by students' characteristics (Doherty & Conolly, 1985; Hoge & Butcher, 1984; Leinhardt, 1983).

We see three main limitations in our study. These are also limitations of the broader literature, so we discuss them in some detail in hopes of informing future research. First, it seems possible that teachers misunderstood what was asked of them in ways that are not immediately obvious. As already stated, teachers either knew or were shown the test on which they were asked to estimate performance, they agreed that the test assessed what they taught, and they were no better at estimating percentages of proficient students or proficiency rates for specific topics and items. Yet, incorporating practice exercises with answers would go even further in ensuring understanding.

Second, teachers may have exerted insufficient effort in estimating their students' test scores. As we demonstrate, teachers' estimations are not pure "noise"; they can be partly explained by the differences between their estimations for low- and high-achieving students and

by students' performance on a test of fluid intelligence. It is still possible, however, that offering teachers incentives to provide their most thoughtful estimates of students' scores may improve their accuracy. As we mentioned, in Bangladesh, we offered teachers an incentive for participation in the survey, but they may provide more thoughtful estimations if they were rewarded based on their accuracy.

Lastly, despite the fact that our instructions were extremely clear, teachers may have stated what they *hoped* students' test scores to be—either due to "motivated reasoning" (Bénabou & Tirole, 2016) or wishful thinking—or what they believed would be *acceptable* test scores—due to "social desirability" (Reisinger, 2022). Yet, if these mechanisms are a major determinant of estimations, it is hard to explain why teachers do not expect *all* students to score higher (i.e., why they do not overestimate the scores of all students by a similar margin). Even their highest-achieving students do not obtain perfect test scores, so there is still considerable margin for teachers to expect them to fare even better than they actually do. It is also hard to explain why estimations would be predicted by students' intelligence. Nevertheless, either providing teachers with anchors based on similar students or separately eliciting the scores they wished their students would obtain and the ones that they expect them to obtain could potentially be helpful in eliciting more accurate estimations.

Ultimately, the frontier question for this literature in LMICs is whether (and if so, how) teachers' misestimations of students' test scores affect classroom instruction. Specifically, we hypothesize that teachers may interact less or engage in different types of interactions with students for whom their estimations are least accurate. We see studies that link teachers' (mis)estimations to the frequency and type of student-teacher interactions as a natural next step of this research agenda.

References

Andrabi, T., Das, J., Khwaja, A. I., Vishwanath, T., & Zajonc, T. (2007). *Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to inform the education policy debate*.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, *131*(3), 1415-1453.

ASER. (2023). *Annual status of education report (ASER) 2022: Provisional*.

Azam, M., & Kingdon, G. G. (2015). Assessing teacher quality in India. *Journal of Development Economics*, *117*, 74-83. https://doi.org/https://doi.org/10.1016/j.jdeveco.2015.07.001

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof to concept to scalable policies: Challenges and solutions, with an application *Journal of Economic Perspectives*, *31*(4), 73-102.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy*, *2*, 1-30. https://doi.org/10.1257/pol.2.1.1

Banerjee, A. V., Banerji, R., Duflo, E., & Walton, M. (2011). *Effective pedagogies and a resistant education system: Experimental evidence on interventions to improve basic skills in rural India*.

Banerjee, A. V., Bhattacharjee, S., Chattopadhyay, R., Ganimian, A. J., Duflo, E., & Spelke, E. (2023). *Street smart or school smart? The arithmetic skills of working children in two Indian cities*. Abdul Latif Jameel Poverty Action Lab (J-PAL).

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*, 1235-1264. https://doi.org/10.1162/qjec.122.3.1235

Banerjee, A. V., & Duflo, E. (2011). Top of the Class. In *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. Public Affairs.

Bau, N., & Das, J. (2017). *The misallocation of pay and productivity in the public sector: Evidence from the labor market for teachers*. The World Bank.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly*, *23*(1), 43–55.

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review*, *40*(1), 23–38.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*(3), 141-164.

Bhattacharjea, S., Wadhwa, W., & Banerji, R. (2011). *Inside primary schools: A study of teaching and learning in rural India*. ASER.

Bold, T., Filmer, D. P., Martin, G., Molina, E., Stacy, B., Rockmore, C., Svensson, J., & Wane, W. (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in Africa. *Journal of Economic Perspectives*, *31*(4), 185-204.

Bruns, B., Filmer, D. P., & Patrinos, H. A. (2011). *Making schools work: New evidence on accountability reforms*. World Bank Publications.

Bruns, B., & Luque, J. (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. The World Bank.

Buhl-Wiggers, J., Kerwin, J. T., Smith, J. A., & Thornton, R. (2019). *Teacher effectiveness in Africa: Longitudinal and causal estimates*. Department of Economics, Copenhagen Business School.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: Teacher and health worker absence in developing countries. *The Journal of Economic Perspectives*, *20*(1), 91-116.

de Barros, A., & Ganimian, A. J. (2023). The foundational math skills of Indian children. *Economics of Education Review*, *92*, 102336.

de Hoyos, R., Djaker, S., Ganimian, A. J., & Holland, P. (2023). *Can performance-management tools and training complement diagnostic feedback? Experimental evidence from public schools in Argentina*. New York University.

de Hoyos, R., Ganimian, A. J., & Holland, P. (2021). Teaching with the test: Experimental evidence on diagnostic feedback and capacity-building for schools in Argentina. *World Bank Economic Review*, *35*(2), 499-520.

de Hoyos, R., García-Moreno, V. A., & Patrinos, H. A. (2017). The impact of an accountability intervention with diagnostic feedback: Evidence from Mexico. *Economics of Education Review*, *58*, 123-140.

Doherty, J., & Conolly, M. (1985). How accurately can primary school teachers predict the scores of their pupils in standardised tests of attainment? A study of some non-cognitive factors that influence specific judgements. *Educational Studies*, *11*(1), 41-60.

Duflo, A., Kiessel, J., & Lucas, A. M. (2020). *Experimental evidence on alternative policies to increase learning at scale*. National Bureau of Economic Research (NBER).

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739-1774. https://doi.org/10.1257/aer.101.5.1739

Farfan Bertran, M. G., Holla, A., & Vakis, R. (2021). *Poor expectations: Experimental evidence on teachers' stereotypes and student assessment*. The World Bank.

Ganimian, A. J., & Djaker, S. (2022). *How can developing countries address heterogeneity in students' preparation for school? A review of the challenge and potential solutions*. Steinhardt School of Culture, Education, and Human Development, New York University.

Ganimian, A. J., Mbiti, I. M., & Mishra, A. (2024). *Teach for science: Experimental evidence on a STEM teaching fellowship in India*. Steinhardt School of Culture, Education, and Human Development, New York University.

Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, *4*(4), 146-168.

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology*, *76*(5), 777.

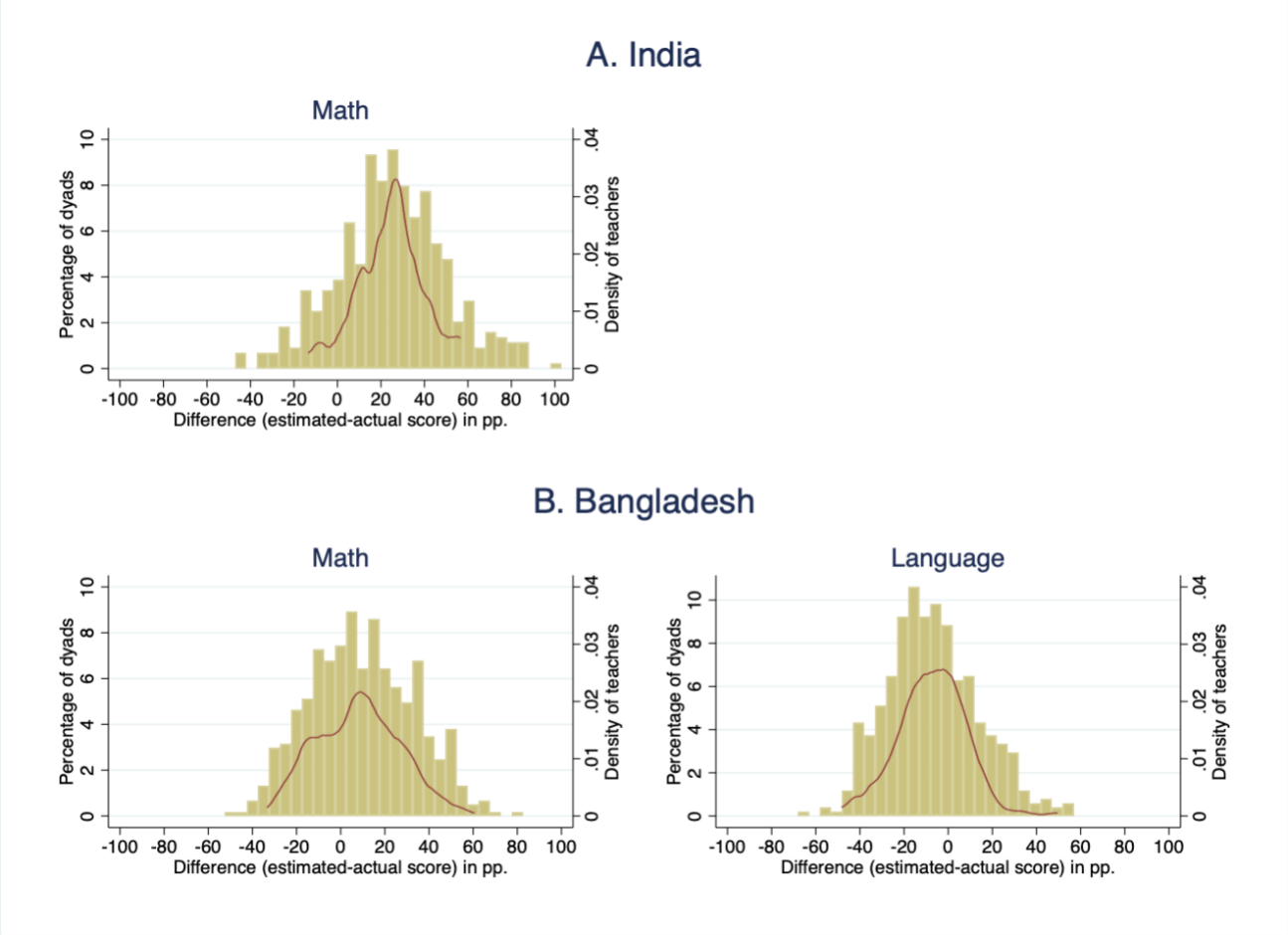
Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297-313.

IEA. (2017). *TIMSS 2019: Assessment frameworks*. TIMSS & PIRLS International Study Center. Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).

Kaufmann, E. (2020). How accurately do teachers' judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology*, *63*, 101902.
Leinhardt, G. (1983). Novice and expert knowledge of individual student's achievement. *Educational Psychologist*, *18*(3), 165-179.
Metzler, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics*, *99*(2), 486-496.
MEW-DSHE. (2019). *Frameworks for National Assessment of Secondary Students 2019. Subject: Bangla, English, and mathematics. Grades: 6, 8, and 10* Monitoring and Evaluation Wing, Directorate of Secondary and Higher Education (MEW-DSHE).
Muralidharan, K., Das, J., Holla, A., & Mohpal, A. (2017). The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics*, *145*(C), 116-135.
Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, *109*(4), 1426-1460.
Pritchett, L., & Beatty, A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development*, *40*, 276-288.
Raven, J., & Summers, W. A. (1986). Research Supplement No. 3.: A Compendium of North American Normative and Validity Studies. In J. C. Raven, J. H. Court, & J. Raven (Eds.), *Manual for Raven's Progressive Matrices and Vocabulary Tests*. Psychological Corporation.
Reisinger, J. (2022). Subjective well-being and social desirability. *Journal of Public Economics*, *214*, 104745.
Sabarwal, S., Abu-Jawdeh, M., & Kapoor, R. (2022). Teacher beliefs: Why they matter and what they are. *The World Bank Research Observer*, *37*(1), 73-106.
Sankar, D., & Linden, T. (2014). *How much and what kind of teaching is there in elementary education in India? Evidence from three states*. The World Bank.
Santibañez, L. (2006). Why we should care if teachers get A's: Teacher test scores and student achievement in Mexico. *Economics of Education Review*, *25*(5), 510-520.
Singh, A. (2019). Learning more with every year: School year productivity and international learning divergence. *Journal of the European Economic Association*, 1-44.
Stallings, J. A., Knight, S. L., & Markham, D. (2014). *Using the Stallings observation system to investigate time on task in four countries*. The World Bank.
Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743.
Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 100374.
Wadmare, P., Nanda, M., Sabates, R., Sunder, N., & Wadhwa, W. (2022). Understanding the accuracy of teachers' perceptions about low achieving learners in primary schools in rural India: An empirical analysis of alignments and misalignments. *International Journal of Educational Research Open*, *3*, 100198.
World Bank. (2018). *World development report 2018: Learning to realize education's promise*. The World Bank.

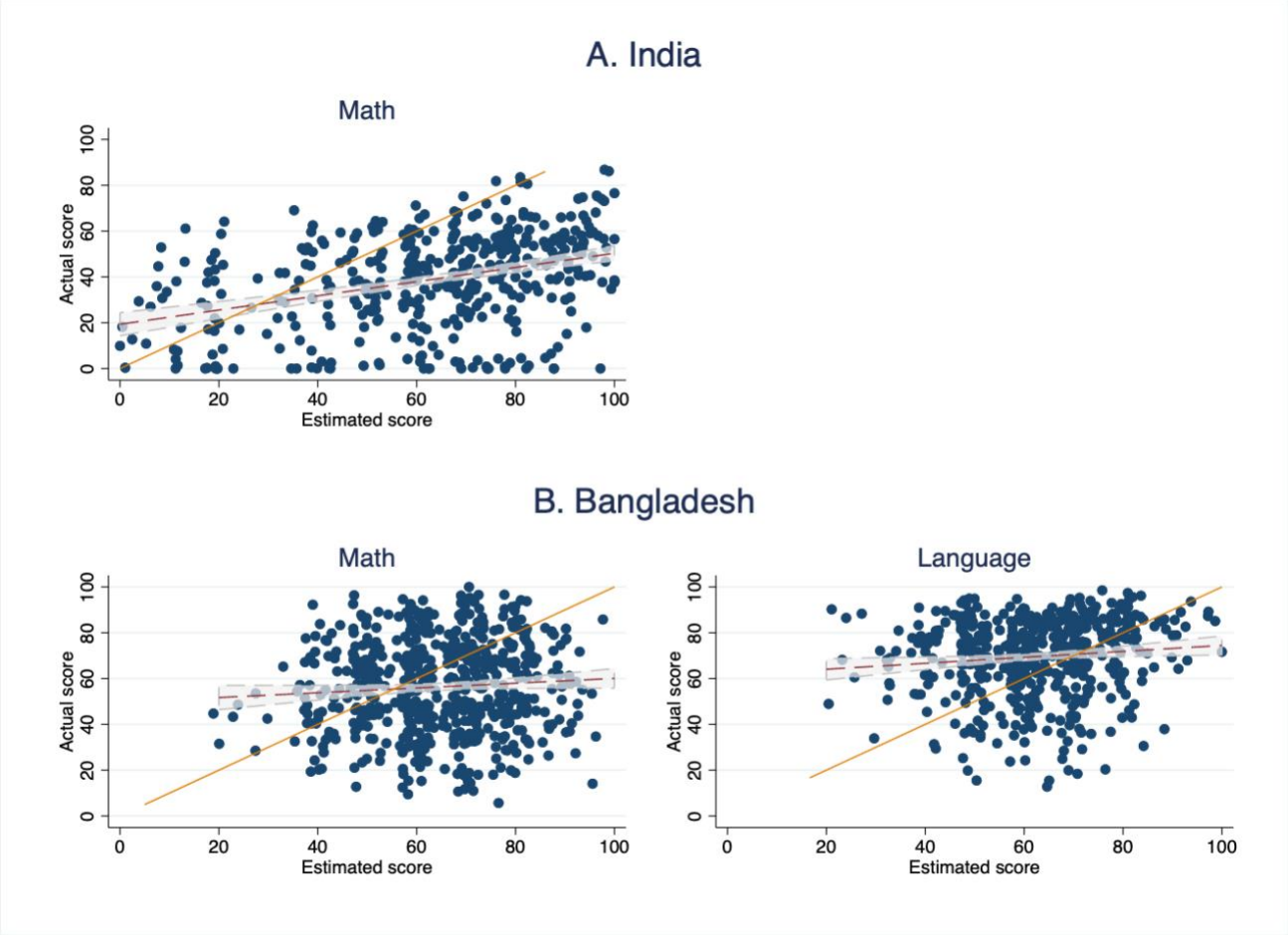
Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)*. American Council on Education and Praeger Publishers.

Figure 1: Distribution of differences between students' test scores and teachers' estimations, by country and subject



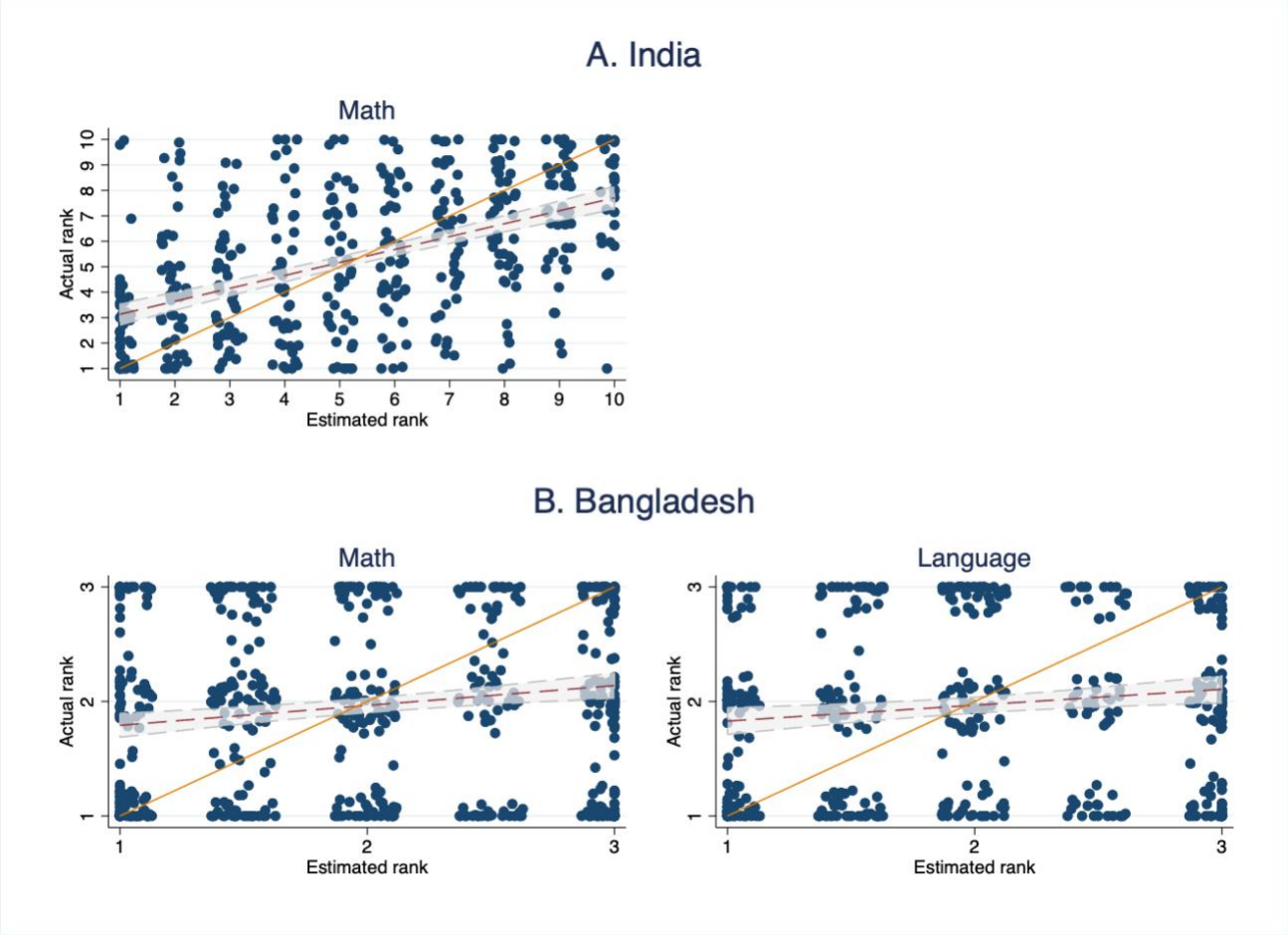
Notes: This graph shows the distribution of differences between students' test scores and teachers' estimations of those scores at the student-teacher dyad level (histogram, left y-axis) and the distribution of average differences at the teacher level (density plot, right y-axis). Panel A shows these distributions for India (where only math was assessed) and Panel B shows them for Bangladesh (where both math and language were assessed). In India, we asked teachers to estimate the scores of 10 students, and in Bangladesh, of three students.

Figure 2: Relationship between students' test scores and teachers' estimations of those scores, by country and subject



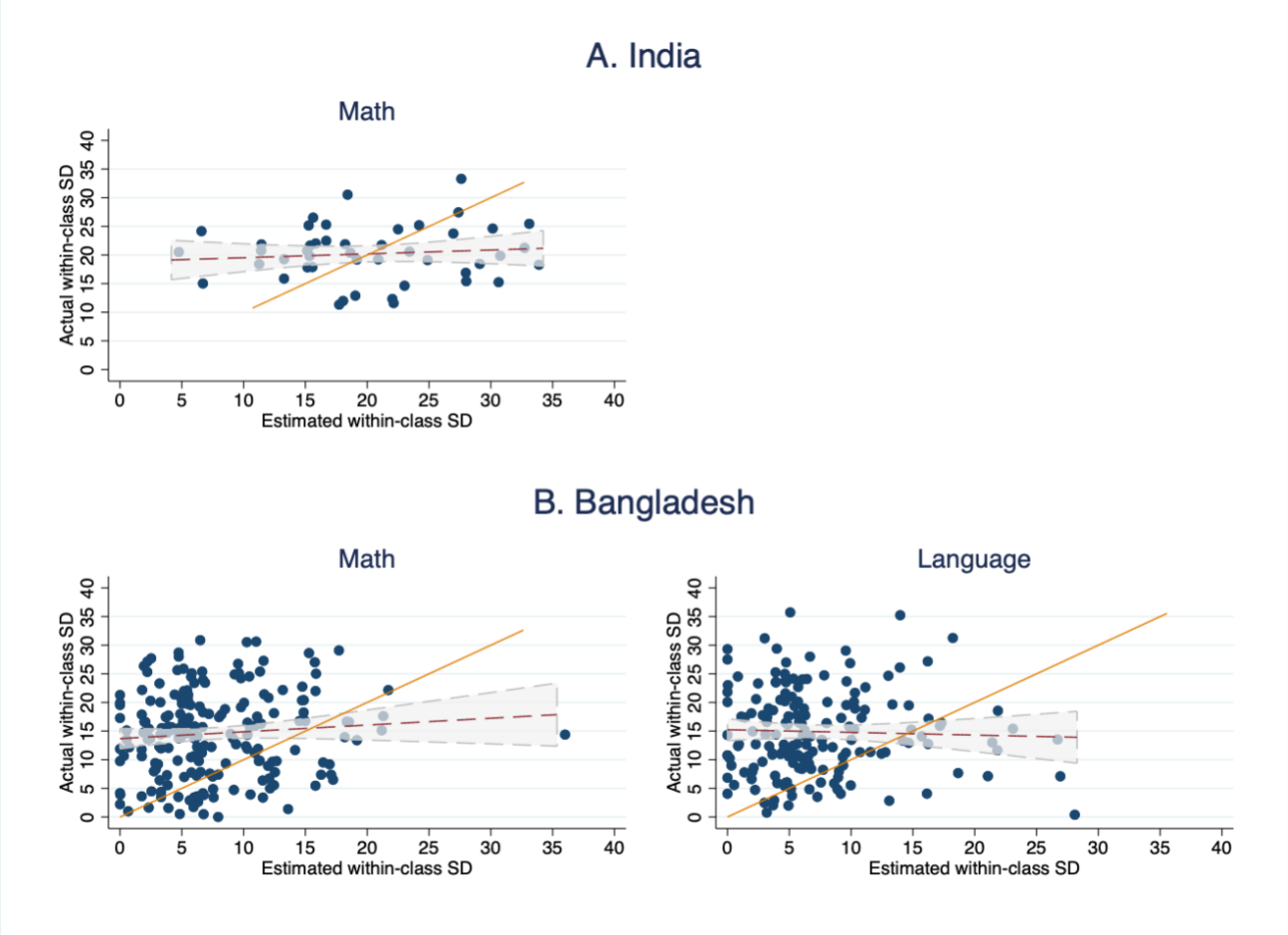
Notes: This graph shows the relationship between students' test scores and teachers' estimations of those scores at the student-teacher dyad level. Panel A shows the scatterplot for India (where only math was assessed) and Panel B the scatterplots for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. The solid 45-degree line indicates where the dots would fall if estimated scores predicted actual scores perfectly. The dashed line is the line of best fit, with the 95% confidence interval shown in gray. We added spherical random noise to the data to show density within each point.

Figure 3: Relationship between students' within-class ranks and their implied within-class ranks in teachers' estimations, by country and subject



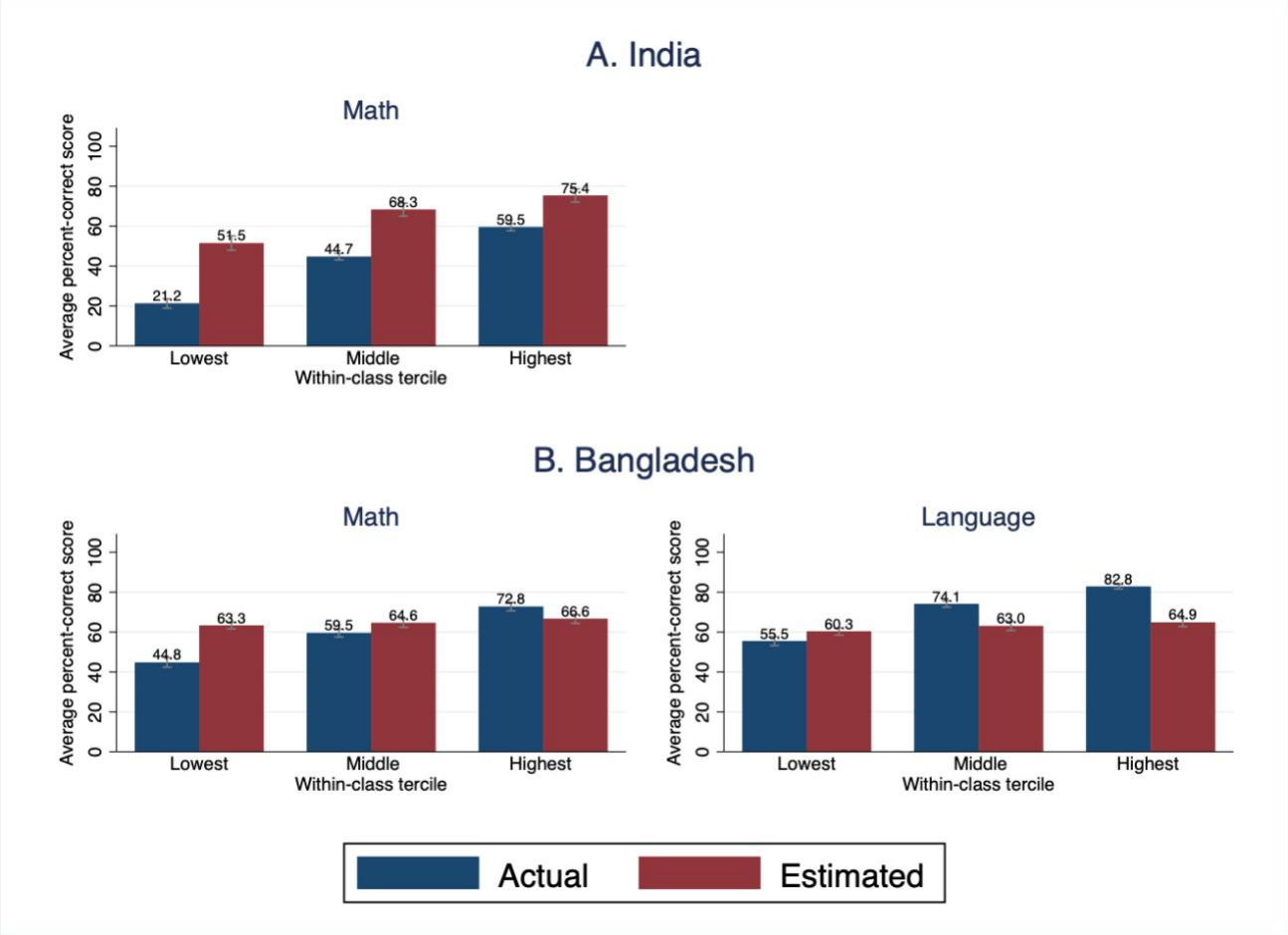
Notes: This graph shows the relationship between students' within-class ranks (based on their actual test scores) and their implied within-class ranks (based on teachers' estimations of their scores) at the student-teacher dyad level. Panel A shows the scatterplot for India (where only math was assessed) and Panel B the scatterplots for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. The solid 45-degree line indicates where the dots would fall if estimated ranks predicted actual ranks perfectly. The dashed line is the line of best fit, with the 95% confidence interval shown in gray. We added spherical random noise to the data to show density within each point.

Figure 4: Relationship between students' actual within-class standard deviations and and their implied within-class standard deviations in teachers' estimations, by country and subject



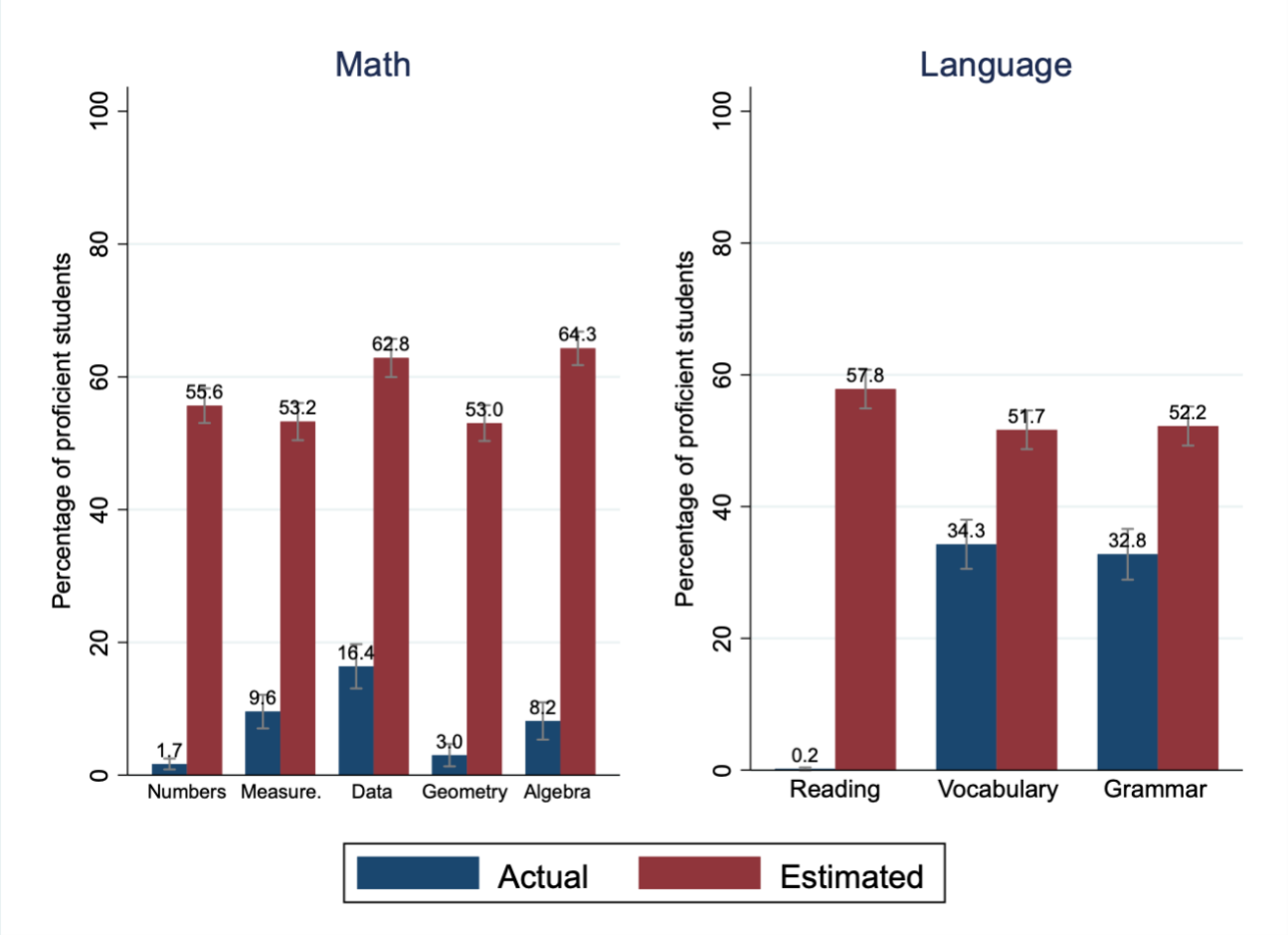
Notes: This graph shows the relationship between students' within-class standard deviations (based on their actual test scores) and their implied within-class standard deviations (based on teachers' estimations of their scores) at the teacher level. Panel A shows the scatterplot for India (where only math was assessed) and Panel B the scatterplots for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. The solid 45-degree line indicates where the dots would fall if estimated standard deviations predicted actual standard deviations perfectly. The dashed line is the line of best fit, with the 95% confidence interval shown in gray. We added spherical random noise to the data to show density within each point.

Figure 5: Students' test scores and teachers' estimations of those scores, by country, subject, and within-class achievement tercile



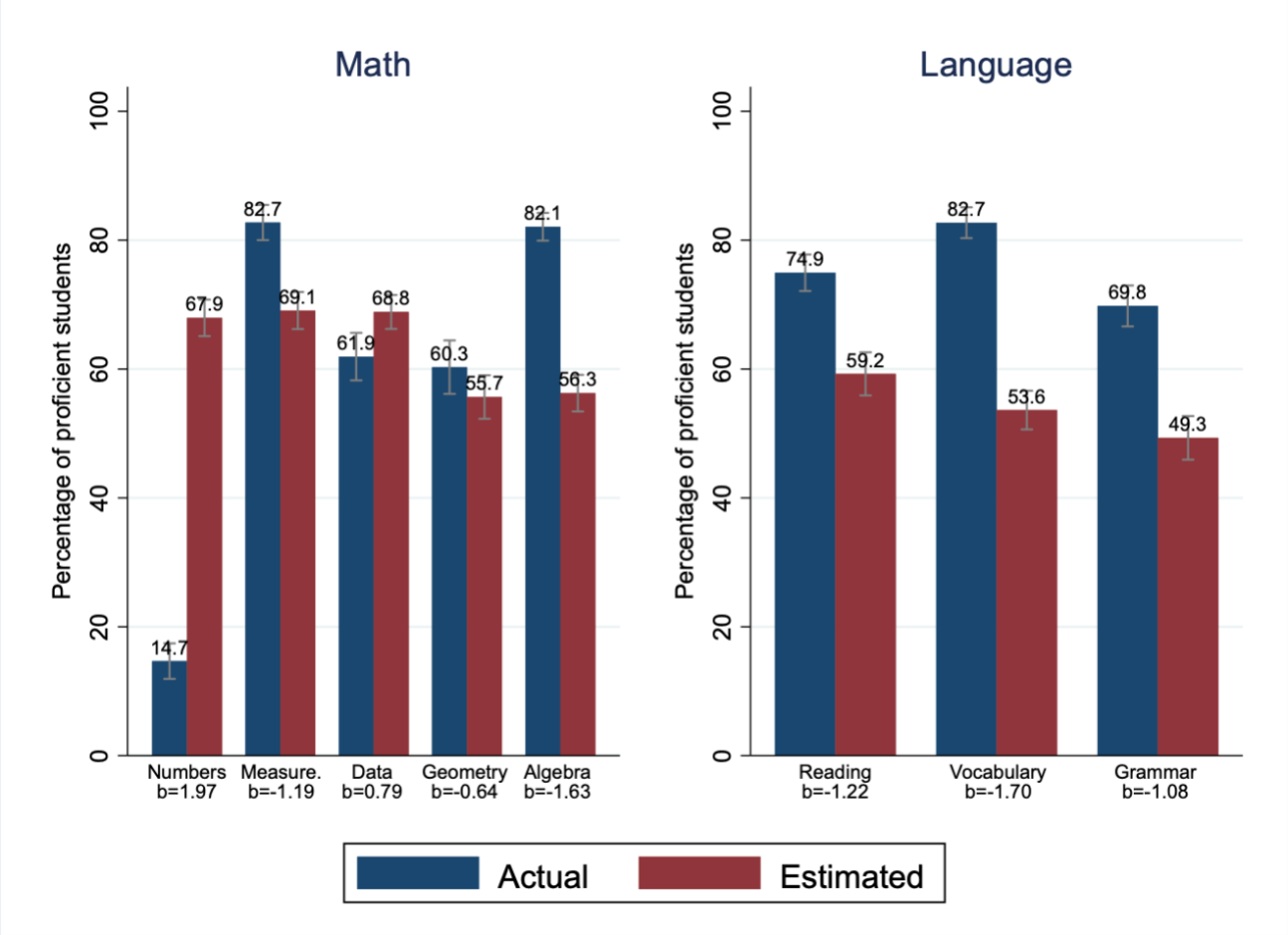
Notes: This graph shows students' test scores and teachers' estimations of those scores at the student-teacher dyad level, by students' within-class achievement tercile. Panel A shows the scatterplot for India (where only math was assessed) and Panel B the scatterplots for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students.

Figure 6: Estimated and actual percentage of proficient students in each topic in Bangladesh, by subject



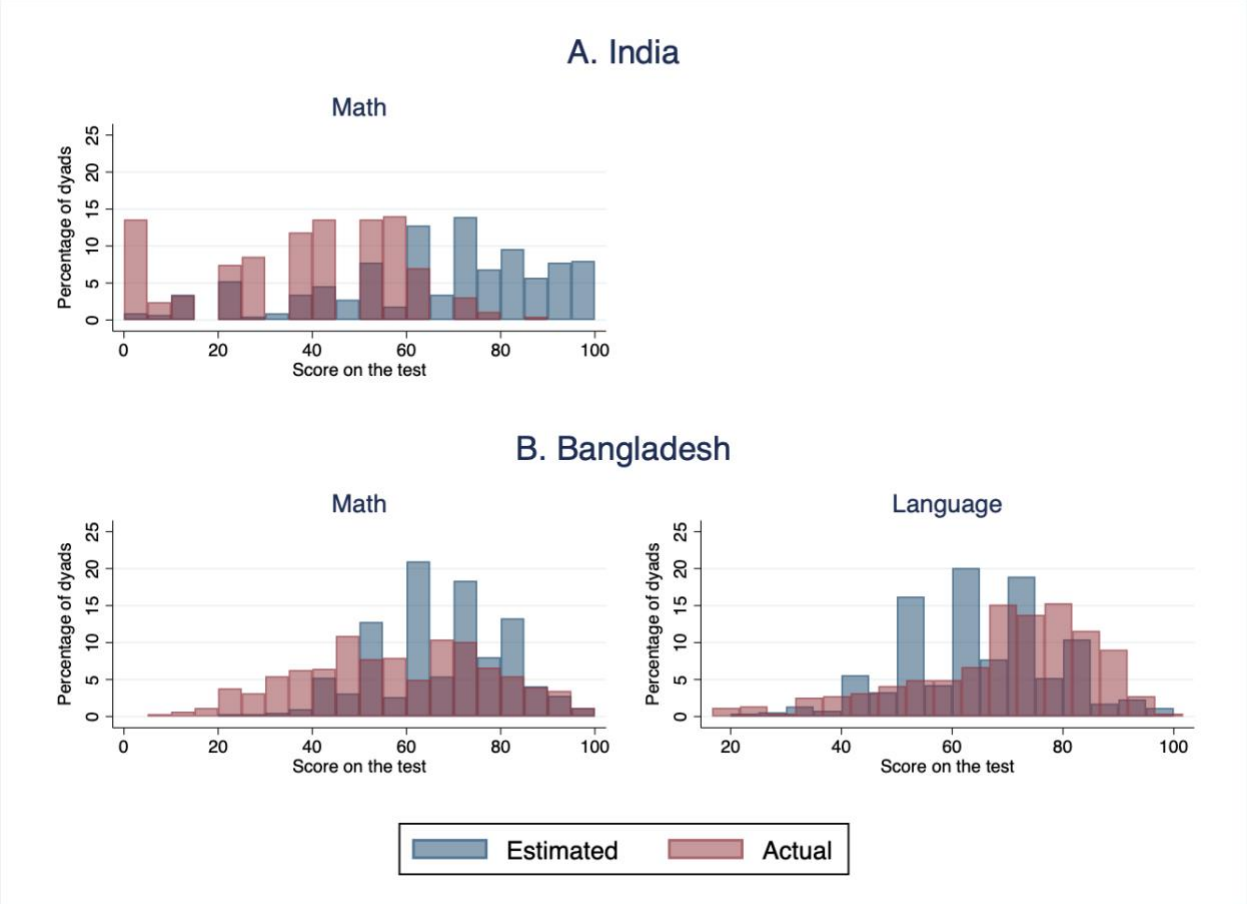
Notes: This graph shows the actual and estimated percentage of "proficient" students on each topic (i.e., students who answer all questions on that topic correctly) based on students' test scores and teachers' estimations, respectively, at the teacher level. We only asked teachers to provide these estimations in Bangladesh. The graph in the left shows results for math and the one in the right for language.

Figure 7: Estimated and actual percentage of proficient students in selected items in Bangladesh, by subject



Notes: This graph shows the actual and estimated percentage of students who can answer specific items on each topic based on students' test scores and teachers' estimations, respectively, at the teacher level. We only asked teachers to provide these estimations in Bangladesh. The graph in the left shows results for math and the one in the right for language. In math, the items included: arranging four numbers based on their value (numbers), calculating the area of a rectangle (measurement), finding the mean and median of five numbers (data), identifying the number of sides in a book (geometry), identifying the coefficient in an algebraic expression (algebra). In language, they included: tracing the cause of an event in a paragraph (reading), finding an antonym (vocabulary), and identifying the components of a word (grammar). The b-parameters indicate the level of latent ability required to answer an item with a 50% chance: negative values indicate easier items and positive values indicate harder items.

Figure A.1: Distribution of students' test scores and teachers' estimations, by country and subject



Notes: This graph shows the distribution of students' test scores and teachers' estimations of those scores at the student-teacher dyad level. Panel A shows these distributions for India (where only math was assessed) and Panel B shows them for Bangladesh (where both math and language were assessed). In India, we asked teachers to estimate the scores of 10 students, and in Bangladesh, of three students.

by country and

Table A.1: Summary statistics for India and Bangladesh

	(1) Mean	(2) SD	(3) Min.	(4) Max.	(5) 25th ptile.	(6) 50th ptile.	(7) 75th ptile.	(8) N (dyads)
A. India								
Math								
Students' actual scores	38.9	20.4	0	86	29	43	57	438
Teachers' estimated scores	63.2	23.7	0	100	50	70	80	438
Estimated-actual (in pp.)	24.3	25.1	-47	100	9	24	40	438
Estimated-actual (as prop. of within-class SD)	1.51	1.06	0	5.90	0.75	1.31	2.04	438
B. Bangladesh								
Math								
Students' actual scores	56.4	19.9	5	100	42.5	57.5	72.5	605
Teachers' estimated scores	64.8	14.0	20	100	55	65	75	605
Estimated-actual (in pp.)	8.44	23.4	-52.5	82.5	-10	7.5	25	605
Estimated-actual (as prop. of within-class SD)	2.18	2.29	0	17.5	0.68	1.55	2.87	601
Language								
Students' actual scores	69.6	16.7	16.7	100	59.5	73.8	80.9	509
Teachers' estimated scores	62.6	13.9	20	100	50	60	70	509
Estimated-actual (in pp.)	-7.02	20.6	-68.1	56.7	-21.0	-8.81	6.19	509
Estimated-actual (as prop. of within-class SD)	1.87	2.29	0	20.5	0.59	1.25	2.26	506

Notes: This table shows summary statistics for the main variables used in both studies, including: students' actual test scores, teachers' estimations of those test scores (both in percentage-correct terms, from 0 to 100), and differences between these two expressed in percentage points and as the proportion of the within-class standard deviation in actual scores at the student-teacher dyad level. In India, we calculate the average difference between scores and estimations as a share of the within-class standard deviation with respect to 10 students about whom teachers were asked to provide estimations. In Bangladesh, we do so with respect to all students who took the national assessment (not just the three students on whom teachers were asked to provide estimations).

Table A.2: Relationship between students' test scores and teachers' estimations of those scores, by country and subject

	(1)	(2)
	Students' actual scores	
A. India		
Math		
Teachers' estimated scores	0.31***	0.45***
	(0.04)	(0.04)
N (dyads)	438	438
R-squared	0.130	0.349
B. Bangladesh		
Math		
Teachers' estimated scores	0.10*	0.32***
	(0.06)	(0.12)
N (dyads)	605	605
R-squared	0.005	0.568
Language		
Teachers' estimated scores	0.13**	0.44***
	(0.05)	(0.11)
N (dyads)	509	509
R-squared	0.012	0.379
Teacher fixed effects?	N	Y

Notes: This table shows the results from ordinary least-squares regressions of students' test scores on teachers' estimations of those test scores at the student-teacher dyad level, without (column 1) and with (column 2) teacher fixed effects. Panel A shows results for India (where only math was assessed) and Panel B for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. Standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Relationship between students' test scores and teachers' estimations of those scores, by subject and teachers' confidence in their estimations, Bangladesh

	(1)	(2)	(3)
		Students' actual scores	
Teacher is…	…certain	…very confident	…somewhat confident
Math			
Teachers' estimated scores	0.32***	0.07	-0.12
	(0.12)	(0.08)	(0.20)
N (dyads)	137	417	51
R-squared	0.051	0.003	0.005
Language			
Teachers' estimated scores	0.25***	0.04	0.14
	(0.08)	(0.09)	(0.12)
N (dyads)	159	266	84
R-squared	0.056	0.001	0.011
Teacher fixed effects?	N	N	N

Notes: This table shows the results from ordinary least-squares regressions of students' test scores on teachers' estimations of those test scores at the student-teacher dyad level by teachers' level of confidence in their estimation, from highest (certain, column 1) to lowest (somewhat confident, column 2). All results are for Bangladesh, the only country in which we asked teachers to report their level of confidence in their estimations. Standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Relationship between students' within-class ranks and their implied within-class ranks in teachers' estimations, by country and subject

	(1)	(2)
	Students' actual ranks	
A. India		
Math		
Teachers' estimated ranks	0.51***	0.51***
	(0.04)	(0.04)
N (dyads)	438	438
R-squared	0.250	0.253
B. Bangladesh		
Math		
Teachers' estimated ranks	-0.17***	0.13*
	(0.05)	(0.07)
N (dyads)	605	605
R-squared	0.025	0.059
Language		
Teachers' estimated ranks	0.14**	0.13*
	(0.06)	(0.08)
N (dyads)	509	509
R-squared	0.016	0.044
Teacher fixed effects?	N	Y

Notes: This table shows the results from ordinary least-squares regressions of students' within-class ranks (based on their actual test scores) on their implied within-class ranks (based on teachers' estimations of their scores) at the student-teacher dyad level, without (column 1) and with (column 2) teacher fixed effects. Panel A shows results for India (where only math was assessed) and Panel B for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. Standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: Relationship between students' actual within-class standard deviations and their implied within-class standard deviations in teachers' estimations, by country and subject

	(1) Students' actual within-class SD
A. India	
Math	
Teachers' estimated within-class SD	0.05
	(0.08)
N (teachers)	46
R-squared	0.011
B. Bangladesh	
Math	
Teachers' estimated within-class SD	0.12
	(0.09)
N (teachers)	202
R-squared	0.007
Language	
Teachers' estimated within-class SD	-0.05
	(0.10)
N (teachers)	173
R-squared	0.001

Notes: This table shows the results from ordinary least-squares regressions of students' within-class standard deviations (based on their actual test scores) on their implied within-class standard deviations (based on teachers' estimations of their scores) at the teacher level. Panel A shows results for India (where only math was assessed) and Panel B for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. Standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6: Students' test scores and teachers' estimations of those scores, by country, subject, and within-class achievement tercile

	(1) Students' actual scores	(2) Teachers' estimated scores	(3) Col. (2)-(1)	(4) Col. (2)-(1)
A. India				
Math				
Students in lowest within-class tercile	21.2	51.5	30.3***	30.3***
	[16.2]	[24.6]	(3.08)	(3.30)
Students in middle within-class tercile	44.7	68.3	23.6***	23.6***
	[9.78]	[19.7]	(2.37)	(2.59)
Students in highest within-class tercile	59.5	75.4	15.9***	15.9****
	[9.69]	[18.1]	(2.37)	(2.65)
N (dyads)	438	438	438	438
B. Bangladesh				
Math				
Students in lowest within-class tercile	44.8	64.3	18.5***	18.6***
	[19.4]	[13.7]	(1.86)	(2.29)
Students in middle within-class tercile	59.6	64.6	5.01***	5.18**
	[13.6]	[14.7]	(1.65)	(2.15)
Students in highest within-class tercile	72.8	66.6	-6.17***	-6.05***
	[12.9]	[13.6]	(1.66)	(2.26)
N (dyads)	567	567	567	567
Language				
Students in lowest within-class tercile	55.5	60.3	4.84***	5.02**
	[15.5]	[13.0]	(1.77)	(2.29)
Students in middle within-class tercile	74.1	63.0	-11.1***	-11.0***
	[10.5]	[14.6]	(1.49)	(1.96)
Students in highest within-class tercile	82.8	64.9	-17.9***	-17.9***
	[6.72]	[12.8]	(1.15)	(1.53)
N (dyads)	477	477	477	477
Teacher fixed effects?			N	Y

Notes: This table shows the mean and standard deviation of students' test scores (column 1) and teachers' estimations of those test scores (column 2) and then estimates the difference between them by students' within-class achievement tercile at the student-teacher dyad level, without (column 3) and with (column 4) teacher fixed effects. Panel A shows results for India (where only math was assessed) and Panel B for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7: Students' test scores and teachers' estimations of those scores in Bangladesh, by subject and national achievement tercile

	(1) Students' actual scores	(2) Teachers' estimated scores	(3) Col. (2)-(1)	(4) Col. (2)-(1)
Math				
Students in lowest national tercile	33.7	64.0	30.4***	30.4***
	[9.37]	[14.2]	(1.43)	(1.79)
Students in middle national tercile	57.7	64.2	6.52***	6.52***
	[7.03]	[14.8]	(1.25)	(1.64)
Students in highest national tercile	79.1	66.0	-13.0***	-13.0***
	[7.69]	[12.7]	(1.22)	(1.57)
N (dyads)	567	567	567	567
Language				
Students in lowest national tercile	48.1	62.1	14.0***	14.0***
	[11.4]	[12.5]	(1.59)	(2.11)
Students in middle national tercile	72.3	61.4	-10.9***	-10.9***
	[4.53]	[13.8]	(1.23)	(1.61)
Students in highest national tercile	85.3	65.2	-20.1***	-20.1***
	[4.13]	[13.9]	(1.18)	(1.57)
N (dyads)	477	477	477	477
Teacher fixed effects?			N	Y

Notes: This table shows the mean and standard deviation of students' test scores (column 1) and teachers' estimations of those test scores (column 2) and then estimates the difference between them by students' national achievement tercile at the student-teacher dyad level, without (column 3) and with (column 4) teacher fixed effects. It only shows results for Bangladesh, where we observe students' performance with respect to all test takers in the country. In Bangladesh, we asked each teacher to estimate the scores of three students. Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.8: Students' math test scores and teachers' estimations of those scores in India, by quartile of students' fluid intelligence

	(1)	(2)	(3)	(4)
	Students' actual scores	Teachers' estimated scores	Col. (2)-(1)	
Students in quartile 1 (lowest)	30.7	51.7	21.0***	21.0***
	[19.3]	[26.0]	(3.72)	(4.09)
Students in quartile 2	38.6	63.2	24.6***	24.6***
	[20.1]	[21.3]	(2.68)	(2.88)
Students in quartile 3	48.9	69.5	20.6***	20.6***
	[16.1]	[22.8]	(3.85)	(4.47)
Students in quartile 4 (highest)	44.8	75.1	30.3***	30.3***
	[20.5]	[17.9]	(2.86)	(3.17)
N (dyads)	438	438	438	438
Teacher fixed effects?			N	Y

Notes: This table shows the mean and standard deviation of students' test scores (column 1) and teachers' estimations of those test scores (column 2) and then estimates the difference between them by the quartile of students' fluid intelligence at the student-teacher dyad level, without (column 3) and with (column 4) teacher fixed effects. It only shows results for India, where we observe students' fluid intelligence. In India, teachers only estimated students' test scores in math. Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.9: Students' math test scores and teachers' estimations of those scores in India, by student and teacher sex

	(1) Students' actual scores	(2) Teachers' estimated scores	(3) Col. (2)-(1)	(4) Col. (2)-(1)
All teachers				
Male students	39.4	61.5	22.1***	22.1***
	[20.1]	[24.8]	(2.89)	(3.04)
Female students	38.7	65.52	26.8***	26.8***
	[20.3]	[22.37]	(2.77)	(2.91)
N (dyads)	430	430	430	430
Male teachers				
Male students	36.0	64.5	28.5***	28.5***
	[15.8]	[19.4]	(3.80)	(3.99)
Female students	36.9	67.6	30.6**	30.6**
	[16.8]	[18.0]	(8.79)	(9.17)
N (dyads)	82	82	82	82
Female teachers				
Male students	40.1	60.8	20.7***	20.7***
	[20.9]	[25.7]	(3.36)	(3.53)
Female students	39.1	65.0	25.9***	25.9***
	[21.1]	[23.4]	(2.69)	(2.82)
N (dyads)	348	348	348	348
Teacher fixed effects?			N	Y

Notes: This table shows the mean and standard deviation of students' test scores (column 1) and teachers' estimations of those test scores (column 2) and then estimates the difference between them by student and teacher sex at the student-teacher dyad level, without (column 3) and with (column 4) teacher fixed effects. It only shows results for India, where we observe both student and teacher sex. In India, teachers only estimated students' test scores in math. Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.10: Students' math test scores and teachers' estimations of those scores in India, by student socio-economic status

	(1)	(2)	(3)	(4)
	Students' actual scores	Teachers' estimated scores	Col. (2)-(1)	
Students in quartile 1 (poorest)	38.6	62.5	23.9***	23.9***
	[21.3]	[25.1]	(3.19)	(3.43)
Students in quartile 2	39.3	67.6	28.3***	28.3***
	[20.5]	[21.5]	(3.14)	(3.56)
Students in quartile 3	39.3	62.1	22.7***	22.7***
	[18.8]	[24.4]	(2.94)	(3.19)
Students in quartile 4 (richest)	39.2	62.8	23.6***	23.6***
	[20.3]	[21.0]	(3.41)	(3.94)
N (dyads)	430	430	430	430
Teacher fixed effects?			N	Y

Notes: This table shows the mean and standard deviation of students' test scores (column 1) and teachers' estimations of those test scores (column 2) and then estimates the difference between them by the quartile of an index of students' household assets at the student-teacher dyad level, without (column 3) and with (column 4) teacher fixed effects. It only shows results for India, where we observe students' socio-economic status through their household assets. In India, teachers only estimated students' test scores in math. Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.11: Students' math test scores and teachers' estimations of those scores in India, by student caste or tribe

	(1) Students' actual scores	(2) Teachers' estimated scores	(3) Col. (2)-(1)	(4) Col. (2)-(1)
Students from scheduled caste	36.7	61.9	25.2***	25.2***
	[20.4]	[23.8]	(2.55)	(2.69)
Students from scheduled tribe	48.0	71.6	23.6*	23.6
	[26.6]	[14.6]	(10.4)	(14.7)
Students from other backward castes	41.9	70.0	28.1***	28.1
	[22.4]	[23.4]	(5.39)	(6.41)
Students from other general categories	39.6	64.1	24.5***	24.5***
	[19.6]	[23.4]	(2.60)	(2.80)
N (dyads)	421	421	421	421
Teacher fixed effects?			N	Y

Notes: This table shows the mean and standard deviation of students' test scores (column 1) and teachers' estimations of those test scores (column 2) and then estimates the difference between them by students' caste or tribe at the student-teacher dyad level, without (column 3) and with (column 4) teacher fixed effects. It only shows results for India, where we observe students' socio-economic status through their household assets. In India, teachers only estimated students' test scores in math. Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.12: Relationship between students' test scores and teachers' estimations of those scores by country, subject, and teacher characteristics

	(1)	(2)	(3)
	Students' actual scores		
	Teachers with MA degree	Teachers with certification	Teachers with above-median experience
A. India			
Math			
Teachers' estimated scores	0.38***	0.37***	0.29***
	(0.05)	(0.07)	(0.05)
Covariate	7.60	6.83	-3.41
	(6.65)	(4.20)	(6.30)
Interaction	-0.17*	-0.03	0.04
	(0.09)	(0.02)	(0.09)
N (dyads)	438	438	438
R-squared	0.146	0.139	0.131
B. Bangladesh			
Math			
Teachers' estimated scores	0.10	0.07	0.22***
	(0.11)	(0.12)	(0.08)
Covariate	-0.09	-1.74	14.9*
	(8.82)	(9.68)	(7.81)
Interaction	<0.01	0.04	-0.22*
	(0.13)	(0.14)	(0.12)
N (dyads)	605	605	605
R-squared	0.005	0.006	0.012
Language			
Teachers' estimated scores	0.12	0.10	0.11
	(0.08)	(0.08)	(0.07)
Covariate	2.27	-1.27	-3.57
	(6.79)	(6.80)	(6.82)
Interaction	0.01	0.05	0.04
	(0.11)	(0.11)	(0.11)
N (dyads)	509	509	509
R-squared	0.019	0.014	0.013

Notes: This table shows the results from ordinary least-squares regressions of students' test scores on teachers' estimations of those test scores, a teacher-level covariate (having a master's degree, a teaching certificate, or above-median teaching experience), and its interaction with estimations at the student-teacher dyad level. Panel A shows results for India (where only math was assessed) and Panel B for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. Standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.13: Percentage of students recognized by teachers in Bangladesh, by whether they are real or fake, subject, and achievement tercile

	(1) Overall	(2) Within-class distribution	(3) National distribution
Math			
All fake students	28.5		
	[45.2]		
All real students	73.6		
	[44.1]		
Students in lowest tercile		74.2	75.2
		[43.8]	[43.2]
Students in middle tercile		71.9	72.6
		[45.0]	[44.6]
Students in highest tercile		74.8	73.2
		[43.5]	[44.3]
N (dyads)	1,689	1,689	1,689
Language			
All fake students	31.1		
	[46.4]		
All real students	76.3		
	[42.5]		
Students in lowest tercile		74.5	76.6
		[43.6]	[42.4]
Students in middle tercile		77.6	75.4
		[41.8]	[43.1]
Students in highest tercile		77.4	77.3
		[41.9]	[41.9]
N (dyads)	1,528	1,528	1,528

Notes: This table shows the percentage of students who were recognized by math and language teachers, depending on whether they were real or fake (column 1), and—among real students—their achievement tercile in the within-class (column 2) or national (column 3) distribution, in Bangladesh. Standard deviations appear in brackets.

Table A.14: Estimated and actual percentage of students in each tercile of the national achievement distribution in Bangladesh, by subject

	(1) Actual percentage of students	(2) Teachers' estimated percentage	(3) Col. (2)-(1)
Math			
Students in lowest national tercile	47.0	27.8	-19.2***
	[30.2]	[18.7]	(2.95)
Students in middle national tercile	41.6	37.3	-4.26**
	[23.1]	[13.9]	(2.13)
Students in highest national tercile	38.7	34.8	-3.86
	[32.0]	[20.1]	(2.80)
N (teachers)	222	222	222
Language			
Students in lowest national tercile	40.9	26.4	-14.5***
	[25.6]	[19.1]	(2.89)
Students in middle national tercile	38.9	35.3	-3.53
	[19.3]	[15.4]	(2.30)
Students in highest national tercile	32.9	38.3	5.43*
	[26.9]	[23.4]	(3.08)
N (teachers)	191	191	191

Notes: This table shows the actual percentage of students in each tercile of the national achievement distribution (column 1) and teachers' estimations of those percentages (column 2) and then estimates the difference between them at the teacher level (column 3). It only shows results for Bangladesh, where we observe actual and estimated percentages of students in each tercile. Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.15: Estimated and actual percentage of proficient students in each topic in Bangladesh, by subject

	(1) Actual percentage of students	(2) Teachers' estimated percentage	(3) Col. (2)-(1)
Math			
Numbers	1.68	55.7	54.0***
	[6.15]	[19.6]	(1.37)
Measurement	9.57	53.3	43.7***
	[19.0]	[21.0]	(1.87)
Data	16.4	62.8	46.4***
	[25.0]	[21.6]	(2.24)
Geometry	3.01	53.0	50.0***
	[12.7]	[20.2]	(1.63)
Algebra	8.16	64.3	56.1***
	[21.1]	[18.9]	(1.94)
N (teachers)	222	222	222
Language			
Reading	0.19	57.8	57.7***
	[1.45]	[20.4]	(1.50)
Vocabulary	34.3	51.7	17.4***
	[25.9]	[20.4]	(2.49)
Grammar	32.8	52.2	19.5***
	[26.8]	[20.6]	(2.36)
N (teachers)	191	191	191

Notes: This table shows the actual percentage of students who are proficient in each subject (column 1) and teachers' estimations of those percentages (column 2) and then estimates the difference between them at the teacher level (column 3). It only shows results for Bangladesh, where we observe actual and estimated percentages by topic. We asked teachers to categorize a student as proficient on a topic if they were "able to answer *all* items on that topic". Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.16: Estimated and actual percentage of students correctly answering selected items in Bangladesh, by subject

	(1) Actual percentage of students	(2) Teachers' estimated percentage	(3) Col. (2)-(1)
Math			
Numbers (*b=1.97*)	14.7	67.9	53.3***
	[21.0]	[21.3]	(2.02)
Measurement (*b=-1.19*)	82.7	69.1	-13.7***
	[20.8]	[21.6]	(2.02)
Data (*b=0.79*)	61.9	68.9	6.92***
	[28.0]	[19.7]	(2.48)
Geometry (*b=-0.64*)	60.3	55.7	-4.63
	[31.5]	[25.3]	(2.83)
Algebra (*b=-1.63*)	82.1	56.3	-25.8***
	[16.4]	[21.3]	(1.86)
N (teachers)	222	222	222
Language			
Reading (*b=-1.22*)	74.9	59.3	-15.7***
	[20.0]	[23.2]	(2.16)
Vocabulary (*b=-1.70*)	82.7	53.6	-29.1***
	[16.8]	[20.8]	(1.94)
Grammar (*b=-1.08*)	69.8	49.3	-20.5***
	[22.4]	[23.5]	(2.21)
N (teachers)	191	191	191

Notes: This table shows the actual percentages of students who can answer specific items on each topic (column 1) and teachers' estimations of those percentages (column 2) and then estimates the difference between them at the teacher level (column 3). It only shows results for Bangladesh, where we observe actual and estimated percentages by item. In math, the items included: arranging four numbers based on their value (numbers), calculating the area of a rectangle (measurement), finding the mean and median of five numbers (data), identifying the coefficient in an algebraic expression (algebra), and identifying the number of sides in a book (geometry). In language, they included: tracing the cause of an event in a paragraph (reading), finding an antonym (vocabulary), and identifying the components of a word (grammar). Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.17: Estimated and actual percentage of students who can correctly answer selected items in Bangladesh, by subject and tercile of the achievement distribution

	(1) Actual percentage of students	(2) Teachers' estimated percentage	(3) Col. (2)-(1)	(4) Actual percentage of students	(5) Teachers' estimated percentage	(6) Col. (5)-(4)
A. Math	Measurement (*b=-1.47*)			Algebra (*b=-1.41*)		
Within-class distribution						
Students in lowest tercile	69.5	84.9	15.4***	72.9	77.4	4.46
	[46.1]	[35.9]	(3.96)	[44.5]	[41.9]	(3.95)
Students in middle tercile	89.5	85.6	-3.99	89.5	80.4	-9.19**
	[30.7]	[32.3]	(3.33)	[30.7]	[39.9]	(4.23)
Students in highest tercile	96.4	90.6	-5.77*	95.6	85.5	-10.1***
	[18.8]	[29.3]	(3.06)	[20.5]	[35.3]	(3.47)
N (dyads)	567	567	567	567	567	567
National distribution						
Students in lowest tercile	56.2	84.5	28.4***	65.5	75.8	10.3***
	[49.7]	[36.3]	(4.76)	[47.7]	[43.0]	(4.80)
Students in middle tercile	91.0	84.5	-6.50**	88.5	80.5	-8.00**
	[28.7]	[36.3]	(3.26)	[32.0]	[39.7]	(4.02)
Students in highest tercile	100	90.8	-9.25***	97.1	86.1	-11.0***
	[0.00]	[29.1]	(2.15)	[16.8]	[34.7]	(2.98)
N (dyads)	567	567	567	567	567	567
B. Language	Reading (*b=-1.29*)			Vocabulary (*b=2.28*)		
Within-class distribution						
Students in lowest tercile	57.4	86.5	29.2***	62.6	91.2	28.6***
	[49.6]	[34.2]	(4.50)	[48.5]	[28.4]	(4.37)
Students in middle tercile	81.3	87.7	6.48	80.6	88.3	7.72*
	[39.2]	[32.9]	(4.21)	[39.7]	[32.2]	(4.17)
Students in highest tercile	94.1	87.5	-6.57*	95.6	94.9	-0.70
	[23.7]	[33.2]	(3.49)	[20.7]	[22.2]	(2.22)
N (dyads)	492	492	492	492	492	492
National distribution						
Students in lowest tercile	46.85	86.0	39.2***	52.5	90.2	37.8***
	[50.1]	[34.8]	(5.21)	[50.1]	[29.8]	(5.35)
Students in middle tercile	81.9	88.1	6.21	81.9	89.3	7.34*
	[38.6]	[32.4]	(4.03)	[38.6]	[31.0]	(3.77)
Students in highest tercile	95.5	87.3	-8.28***	97.5	93.6	-3.82
	[20.7]	[33.5]	(3.16)	[15.8]	[24.5]	(2.38)
N (dyads)	477	477	477	477	477	477

Notes: This table shows the actual percentages of specific students who can answer specific items (columns 1 and 4) and teachers' estimations of those percentages (columns 2 and 5) and then estimates the difference between them at the student-teacher dyad level (columns 3 and 6). It only shows results for Bangladesh, where we observe actual and estimated percentages by item. In math, the items included: calculating the area of a rectangle (measurement) and identifying the coefficient in an algebraic expression (algebra). In language, they included: identifying the main theme in a passage (reading) and finding the word for an animal that can live in land and water (vocabulary). Standard deviations appear in brackets and standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.18: Relationship between students' test scores and teachers' estimations of those scores omitting estimations at or around 60%, by country and subject

	(1)	(2)	(3)	(4)
	Students' actual scores			
	Omitting estimations exactly at 60%		Omitting estimations within 5 pp. of 60%	
A. India				
Math				
Teachers' estimated scores	0.31***	0.44***	0.31***	0.45***
	(0.04)	(0.04)	(0.04)	(0.04)
N (dyads)	388	388	362	362
R-squared	0.146	0.363	0.153	0.379
B. Bangladesh				
Math				
Teachers' estimated scores	0.09	0.36***	0.09	0.38***
	(0.06)	(0.14)	(0.06)	(0.14)
N (dyads)	477	477	429	429
R-squared	0.005	0.647	0.006	0.671
Language				
Teachers' estimated scores	0.12**	0.49***	0.11**	0.51***
	(0.05)	(0.13)	(0.05)	(0.16)
N (dyads)	412	412	348	348
R-squared	0.012	0.455	0.013	0.517
Teacher fixed effects?	N	Y	N	Y

Notes: This table shows the results from ordinary least-squares regressions of students' test scores on teachers' estimations of those test scores at the student-teacher dyad level omitting estimations at or within 5 pp. of 60%, without (columns 1 and 3) and with (columns 2 and 4) teacher fixed effects. Panel A shows results for India (where only math was assessed) and Panel B for Bangladesh (where both math and language were assessed). In India, we asked each teacher to estimate the scores of 10 students, and in Bangladesh, of three students. Standard errors (clustered by teacher) appear in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.19: Overview of main results, by country and subject

	(1) Math	(2) Math	(3) Language
A. The Magnitude of Teachers' Errors is Large			
Estimated-actual scores (in pp.)	24.3	8.44	-7.02
Estimated-actual scores (as prop. of within-class SD)	1.51	2.18	1.87
Pearson correlation between estimated and actual scores	0.36	0.07	0.11
OLS coeff. of actual on estimated scores	0.31	0.10	0.13
OLS coeff. of actual on estimated scores (with teacher FEs)	0.45	0.32	0.44
B. Teachers Do Not Know They Are Making Mistakes			
Pct. of teachers "certain" or "very confident" on estimations	-	91	83
OLS coeff. of actual on estimated scores ("certain" only)	-	0.32	0.25
OLS coeff. of actual on estimated scores ("v. confident" only)	-	0.07	0.04
OLS coeff. of actual on estimated scores ("s. confident" only)	-	-0.12	0.14
C. Teachers Do Not Know How Their Students Compare			
Spearman correlation between estimated and actual scores	0.35	0.07	0.11
OLS coeff. of actual on estimated ranks	0.51	-0.17	0.14
OLS coeff. of actual on estimated ranks (with teacher FEs)	0.51	0.13	0.13
Pct. of teachers with "flipped" rankings	-	11	12
D. Teachers Underestimate How Much Their Students Vary			
Pct. of teachers who underestimate within-class SD	54	77	78
Pearson correlation between estimated and actual within-class SDs	0.08	0.08	-0.03
OLS coeff. of actual on estimated within-class SD	0.05	0.12	-0.05
E. Teachers Overestimate the Scores of Low Achievers			
OLS coeff. of actual on estimated scores (lowest within-class tercile)	30.3	18.5	4.84
OLS coeff. of actual on estimated scores (middle within-class tercile)	23.6	5.01	-11.1
OLS coeff. of actual on estimated scores (highest within-class tercile)	15.9	-6.17	-17.9
OLS coeff. of actual on estimated scores (lowest national tercile)	-	30.4	14.0
OLS coeff. of actual on estimated scores (middle national tercile)	-	6.52	-10.9
OLS coeff. of actual on estimated scores (highest national tercile)	-	-13.0	-20.1
F. Teachers Overweight the Importance of Intelligence for Test Scores			
OLS coeff. of actual on estimated scores (lowest quartile, fluid intelligence)	21.0	-	-
OLS coeff. of actual on estimated scores (second quartile, fluid intelligence)	24.6	-	-
OLS coeff. of actual on estimated scores (third quartile, fluid intelligence)	20.6	-	-
OLS coeff. of actual on estimated scores (highest tercile, fluid intelligence)	30.3	-	-
G. Teachers Overestimate Girls' Scores by a Larger Amount Than Boys' Scores			
OLS coeff. of actual on estimated scores (all teachers, male students)	22.1	-	-
OLS coeff. of actual on estimated scores (all teachers, female students)	26.8	-	-
OLS coeff. of actual on estimated scores (male teachers, male students)	28.5	-	-
OLS coeff. of actual on estimated scores (male teachers, female students)	30.6	-	-
OLS coeff. of actual on estimated scores (female teachers, male students)	20.7	-	-
OLS coeff. of actual on estimated scores (female teachers, female students)	25.9	-	-
H. Teachers Overestimate Scores of Low- and High-Income Students Similarly			
OLS coeff. of actual on estimated scores (lowest quartile, SES)	23.9	-	-
OLS coeff. of actual on estimated scores (second quartile, SES)	28.3	-	-
OLS coeff. of actual on estimated scores (third quartile, SES)	22.7	-	-
OLS coeff. of actual on estimated scores (highest tercile, SES)	23.6	-	-
I. Teachers Overestimate Scores of Low- and High-Income Students Similarly			
OLS coeff. of actual on estimated scores (scheduled caste)	25.2	-	-
OLS coeff. of actual on estimated scores (scheduled tribe)	23.6	-	-
OLS coeff. of actual on estimated scores (other backward castes)	28.1	-	-
OLS coeff. of actual on estimated scores (other general categories)	24.5	-	-
J. Teachers with More Education, Training, and Experience Are No More Accurate			
Interaction mean in OLS of actual on estimated scores (teachers with MA degree)	0.21	0.10	0.13
Interaction mean in OLS of actual on estimated scores (teachers with certification)	0.34	0.11	0.15
Interaction mean in OLS of actual on estimated scores (teachers with MA degree)	0.33	0	0.15

Notes: This table offers an overview of all the main results presented in the paper. For panels E-I, we omit estimates with teacher fixed effects because they are very close to those without (see Tables A.6-A.11 for full results). SD refers to standard deviation. FEs refer to fixed effects. SES refers to socio-economic status. Interaction mean refers to the sum of the first and third coefficients in each regression of Table A.12.

Table A.20: Overview of robustness checks, Bangladesh, by subject

	(1) Math	(2) Language
A. Errors Are Not Due to Focus on Non-Tested Skills		
Pct. of teachers who believe test assesses what they are teaching (to "high" or "moderate" extent)	96	96
B. Errors Are Not Due to Teachers Not Recognizing Students		
Pct. of students recognized (all fake students)	28.5	31.1
Pct. of students recognized (all real students)	73.6	76.3
Pct. of students recognized (lowest within-class tercile)	74.2	74.5
Pct. of students recognized (middle within-class tercile)	71.9	77.6
Pct. of students recognized (highest within-class tercile)	74.8	77.4
Pct. of students recognized (lowest national tercile)	75.2	76.6
Pct. of students recognized (middle national tercile)	72.6	75.4
Pct. of students recognized (highest national tercile)	73.2	77.3
C. Errors Are Not Due to the Challenge of Estimating Test Scores		
OLS coeff. of actual on estimated pct. students in lowest national tercile	-19.2	-14.5
OLS coeff. of actual on estimated pct. students in lowest national tercile	-4.26	-3.53
OLS coeff. of actual on estimated pct. students in lowest national tercile	-3.86	5.43
Pct. of teachers "surprised" or "very surprised" by pct. of students in each national tercile	63	60
D. Errors Are Not Due to the Difficulty of Aggregating Item Performance		
OLS coeff. of actual on estimated pct. of proficient students in numbers	54.0	-
OLS coeff. of actual on estimated pct. of proficient students in measurement	43.7	-
OLS coeff. of actual on estimated pct. of proficient students in data	46.4	-
OLS coeff. of actual on estimated pct. of proficient students in geometry	50.0	-
OLS coeff. of actual on estimated pct. of proficient students in algebra	56.1	-
OLS coeff. of actual on estimated pct. of proficient students in reading	-	57.7
OLS coeff. of actual on estimated pct. of proficient students in vocabulary	-	17.4
OLS coeff. of actual on estimated pct. of proficient students in grammar	-	19.5
OLS coeff. of actual on estimated pct. answering item in numbers	53.5	-
OLS coeff. of actual on estimated pct. answering item in measurement	-13.7	-
OLS coeff. of actual on estimated pct. answering item in data	6.92	-
OLS coeff. of actual on estimated pct. answering item in geometry	-4.63	-
OLS coeff. of actual on estimated pct. answering item in algebra	-25.8	-
OLS coeff. of actual on estimated pct. answering item in reading	-	-15.7
OLS coeff. of actual on estimated pct. answering item in vocabulary	-	-29.1
OLS coeff. of actual on estimated pct. answering item in grammar	-	-20.5
OLS coeff. of actual on estimated pct. answering item in measurement (lowest within-class tercile)	15.4	-
OLS coeff. of actual on estimated pct. answering item in measurement (middle within-class tercile)	-3.99	-
OLS coeff. of actual on estimated pct. answering item in measurement (highest within-class tercile)	-5.77	-
OLS coeff. of actual on estimated pct. answering item in measurement (lowest national tercile)	28.4	-
OLS coeff. of actual on estimated pct. answering item in measurement (middle national tercile)	-6.50	-
OLS coeff. of actual on estimated pct. answering item in measurement (highest national tercile)	-9.25	-
OLS coeff. of actual on estimated pct. answering item in algebra (lowest within-class tercile)	4.46	-
OLS coeff. of actual on estimated pct. answering item in algebra (middle within-class tercile)	-9.19	-
OLS coeff. of actual on estimated pct. answering item in algebra (highest within-class tercile)	-10.1	-
OLS coeff. of actual on estimated pct. answering item in algebra (lowest national tercile)	10.3	-
OLS coeff. of actual on estimated pct. answering item in algebra (middle national tercile)	-8.00	-
OLS coeff. of actual on estimated pct. answering item in algebra (highest national tercile)	-11.0	-
E. Errors Are Not Due to Overreliance on Heuristics		
Pct. of estimations of students' percent-correct scores exactly at 60%	21	19
Pct. of estimations of students' percent-correct scores within 5 pp. of 60%	29	32
OLS coeff. of actual on estimated scores (omitting estimations exactly at 60%)	0.09	0.09
OLS coeff. of actual on estimated scores (omitting estimations exactly at 60%, with teacher FEs)	0.36	0.38
OLS coeff. of actual on estimated scores (omitting estimations within 5 pp. of 60%)	0.12	0.11
OLS coeff. of actual on estimated scores (omitting estimations within 5 pp. of 60%, with teacher FEs)	0.49	0.51

Notes: This table offers an overview of all the robustness checks presented in the paper. We omit item-by-student-tercile results for language from Table A.17, which consistent with those for math but are too many to include here. We also omit results for Tables A.18 for India, which are the only set of robustness checks in the paper for that country. FEs refer to fixed effects.