Policy Research Working Paper 4644

India Shining *and* Bharat Drowning:

Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement

Jishnu Das
Tristan Zajonc

The World Bank
Development Research Group
Human Development and Public Services Team
June 2008

Policy Research Working Paper 4644

Abstract

This paper uses student answers to publicly released questions from an international testing agency together with statistical methods from Item Response Theory to place secondary students from two Indian states—Orissa and Rajasthan—on a worldwide distribution of mathematics achievement. These two states fall below 43 of the 51 countries for which data exist. The bottom 5 percent of children rank higher than the bottom 5 percent in only three countries—South Africa, Ghana and Saudi Arabia. But not all students test poorly. Inequality in the test-score distribution for both states is next only to South Africa in the worldwide ranking exercise. Consequently, and to the extent that these two states can represent India, the two statements ``for every ten top performers in the United States there are four in India'' and ``for every ten low performers in the United States there are two hundred in India'' are both consistent with the data. The combination of India's size and large variance in achievement give both the perceptions that India is shining even as Bharat, the vernacular for India, is drowning. Comparable estimates of inequalities in learning are the building blocks for substantive research on the correlates of earnings inequality in India and other low-income countries; the methods proposed here allow for independent testing exercises to build up such data by linking scores to internationally comparable tests.

This paper—a product of the Human Development and Public Services Team, Development Research Group—is part of a larger effort in the department to measure and understand inequality in the provision of education. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at jdas1@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

India Shining *and* Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement[*]

Jishnu Das
Center for Policy Research, New Delhi
World Bank, Washington DC

Tristan Zajonc
Harvard University

"It has long been claimed that everything one can say about India is true—and so is the opposite." -Devesh Kapur, 2005

1 Introduction

Net primary enrollment in India has risen steadily over the last several decades and now exceeds 90 percent in most of the country. Large planned increases in the government education budget suggest renewed interest and action on the part of the state, with an emphasis on secondary schooling. Not surprisingly, increasing enrollments and resources have shifted the debate from how many children are in school to what they are learning. A consensus is building that getting children into schools may not be enough. Filmer et al. (2006) go so far as to propose augmenting the Millennium Development Goals with a Millennium Learning Goal that provides international benchmarks on how much children know at a pre-specified age. We ask the following question: Is there a way to place Indian children in secondary schools on an international scale (given India's

[*] We thank Lant Pritchett for extensive discussions of the paper. Kin Bing Wu who led a World Bank sector study on secondary education in India designed the collection of the data we use here, and we are grateful to her for making the data and her report available to us. Eric Hanushek and Eugene Gonzalez provided invaluable comments on an early version of this paper and their insights have been critical for the current revision. Michelle Riboud and Sam Carlson provided useful comments that pertain to India's education sector. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent. Working papers describe research in progress by the authors and are published to elicit comments and to further debate.

reluctance to participate in internationally benchmarked tests) and, if so, what would we find in terms of the average score and variance of the achievement distribution?

We propose a method that uses publicly released questions (items) from the Trends in International Mathematics and Science Study (TIMSS) 1999 8th-Grade Mathematics test to place Indian students on an internationally comparable achievement scale. The test, which consists of 36 items taken from the full TIMSS item bank, was administered to 6,000 students in public and private schools in two Indian states—Rajasthan and Orissa. Using the published item parameters for these 36 questions in conjunction with the Item Response Theory test-equating methods used by TIMSS, we construct a distribution of scores for the tested children that is directly comparable to the worldwide distribution; this allows us to compare the tested children to the international average and to place them in reference to the 51 other countries tested by TIMSS in 1999 and 2003.

The average scores of children in Rajasthan and Orissa place these states below 46 and 42 of the 51 countries tested in 1999 or 2003. After nine years of education, between 30 and 40 percent of enrolled children in these two states cannot pass a low international benchmark, described as "some basic mathematical knowledge." Children enrolled in secondary schools in these two Indian states are 3.1 (OECD) standard-deviations below the OECD mean. Where children in these two states are relative to the rest of the world is harder to ascertain. On the one hand, the TIMSS sample is heavily biased towards relatively high-income countries. The median scores, for instance, in Rajasthan and Orissa do not look too bad compared to Philippines and Chile. On the other hand, secondary school enrollments in India are also lower—53 percent of the appropriate age group is enrolled, compared to more than 90 percent in South Africa, the worst performer in the TIMSS sample. To the extent that children currently out of school are less "motivated" or "able", test scores would arguably look worse at higher levels of enrollment.

The test-score distribution is also highly unequal—the difference between the top 5 percent and bottom 5 percent in both states is among the highest in the world, next only to South Africa. Students at the bottom of the distribution in both states score similarly or worse than the bottom students in the three worst performing countries. At the same time, students at the top of the distribution score higher than the top students in other low performing countries, and higher than the median student in all but the best countries. The top 5 percent of students in

Orissa, for example, score higher than the median student in more than 42 of 46 countries tested in 2003.

Faced with similar results on learning, defenders of the quality of education in Indian schools often point to the large number of globally competitive Indians. We perform the following thought experiment: Suppose that these two states represent India (more on this below). Could the country's size combined with the large variance in scores explain how divergent beliefs can be sustained by the same data? As it turns out, in absolute terms, India has just under half the number of 14-year olds who pass the advanced international benchmark as the United States—100 thousand compared to 250 thousand—and roughly the same number who pass the intermediate international benchmark. Indeed, India has more top achievers than any European country tested, which, although not surprising given India's size, helps explain India's visible position on the academic stage. But another view is also sustainable. The average child scores far below any reasonable curricular standard and a large minority in these two states fails completely. If the results form these two states hold more generally, over 18 million 14-year olds in India are either not enrolled or are failing the lowest international benchmark if enrolled. That number is 22 times the number of failing children in the United States and more than any other country tested.

Beyond providing illustrative results for India, this paper is about the building blocks for research on learning and learning inequality in low-income countries where data on internationally comparable tests are typically absent. This requires 1) techniques to place individual students on a single comparable achievement metric and 2) methods to calculate other population quantities, such as the fraction of children passing particular criterion-referenced thresholds or the 5th to 95th percentile achievement spread. Clarifying what is required for comparable measures of learning and learning dispersion allows the research to focus on substantive rather than statistical issues, without worrying about whether results are driven by measurement tools and differing methodologies.

To preview the methodology, independent tests can be linked to the TIMSS achievement distribution provided at least one question is drawn from the TIMSS item bank to fix the free parameters. The primary methodological difficulty arises because "knowledge" or "achievement" is inferred from the data rather than directly observed. Since individual knowledge is measured with error, the variance of the achievement distribution aggregated from Maximum Likelihood estimates of individual knowledge overestimates the true variance. An alternate method, outlined by Mislevy, Beaton, Kaplan & Sheehan (1992), draws from the posterior of every student's achievement distribution to obtain an unbiased measure of the full learning distribution. These draws—known as "plausible values"—are interpreted as individual achievement with the property that when aggregated to a population distribution they recover the correct population moments. We show that the variance of the distribution is sensitive to the estimation method used (i.e. Maximum Likelihood, Bayesian, or Plausible Values), primarily because the TIMSS test is too difficult for a large fraction of Indian children.[1] The method of plausible values offers an alternative for the calculation of higher moments in any setting—such as poverty mapping—where individuals attributes are estimated with a known standard error.

Linking scores to an international distribution contributes to the literature on education in low-income countries in several ways. First, linked test scores are comparable across space and time. Despite increasing worldwide testing using standardized methods—e.g. TIMSS (51 countries), PIRLS (35 countries), IALS (22 countries) and PISA (49 countries)—the Indian government, like many others, is reluctant to participate in such large-scale testing exercises. As a result, what little is known about learning achievement in India, and most low-income countries, arises from an ad-hoc collection of criterion-referenced exams.[2] These tests, administered by independent agencies, are typically not validated using standard testing tools, cannot be equated over time or across countries, and are not subject to a battery of robustness checks that accompany large-scale testing in the OECD countries. The methods applied here allow independent researchers to report achievement distributions for the tests they control that are directly

[1] Brown & Micklewright (2004) also highlights the importance of using a consistent methodology. They find, for instance, that rankings of countries by within-country difference in TIMSS changed substantially for some countries when the scoring model used in 1999 was retrospectively applied to 1995 data.

[2] Examples for India include a large national study by the National Center for Educational Research and Training (NCERT) in 1994, which found that children scored an average of 47 percent in language and 41 percent in mathematics (Shukla et al. 1994), and state-wide studies with smaller samples in Bihar, Tamil Nadu, Delhi, Uttar Pradesh, Madhya Pradesh and Rajasthan(Bashir 1994, Hasan 1995, Govinda & Varghese 1993, Aggarwal 2000, Goyal 2007). In a major recent effort, the NGO Pratham tested children from almost all districts and found low levels of learning: 52 percent of children between the ages of 7 and 10 could read a small paragraph with short sentences at first grade difficulty levels, 32 percent could read a story text and 54 per cent were unable to divide or subtract (Pratham 2006). Similar results have been reported for Africa. In a relatively large effort, the Monitoring Learning Achievement Project (Chinapah et al. 2000, Strauss & Burger 2000)covered 13 African countries and found literacy, numeracy, and life-skills scores for fourth graders between 30 and 70 percent.

comparable to those obtained worldwide.[3]

Comparable achievement measures contribute to our understanding of earnings inequality and its correlates. A growing literature examines the relationship between earnings inequality and test-score dispersion. Nickell (2004) and Blau & Kahn (2005) report a high correlation between test-score dispersion and wage inequality; Nickell (2004) for instance suggests that 70 percent of the dispersion in earnings internationally is attributable to the dispersion in test-scores. Similarly, Bedard & Ferrall (2003) show that test-score inequality at early ages is correlated with wage inequality in the same cohort later in life. In contrast to this literature, Devroye & Freeman (2001) argue that wage dispersion within narrowly defined skill sets is higher than that across and that institutional mechanisms of collective bargaining matter more. India has recently seen a dramatic increase in inequality (Debroy & Bhandari 2007), at the same time that inequality in educational attainment is falling (Jalan & Murgai 2007). It is likely that as inequality in attainment declines further and returns to skill increase (Kijima 2006), attention will increasingly focus on the inequality in cognitive ability.

The remainder of this paper is structured as follows. Section 2 outlines the Item Response Theory method for equating test scores. The technical section and accompanying appendix provides sufficient details for critique and replication. Section 3 discusses the data, sampling strategy, and test design. Section 4 reports the international benchmarking results and variance decompositions. Section 5 outlines some caveats to our method and several robustness checks; Section 6 concludes.

2 Overview of Linking Methodology

Properly linking India's mathematics achievement to the world distribution requires either a *single* test given across all countries (and each year) or a means of linking alternate test forms which may include different items. Since giving a single test is clearly infeasible in most situations, educational testing organizations have developed statistical tools that allow scores from different exams to be expressed on a unified scale. Item Response Theory (IRT) is one such technique

[3] Inter alia, such standardization would help benchmark the relative efficacy of different educational interventions. High inequality in the Indian learning distribution, for instance, implies that interventions leading to a, say, 0.2 s.d. increase in learning represent a greater increase in "knowledge" than a similar effect in the United States.

and is used in most large-scale testing situations such as TIMSS, PIRLS, NAEP and the SAT and GRE. The basic intuition behind this technique is to model the behavior of each item—i.e. its difficulty, ability to discriminate between two children, and likelihood of being guessed—so that any differences in items can be removed from the score. This contrasts with the commonly reported percent correct score, which gives performance on a test-specific scale.

The fundamental building block of IRT is therefore the *item response function* (IRF), which links the latent ability, θ, to the probability a randomly drawn examinee of a given ability will answer the item correctly. One of the most popular models for dichotomous responses is the three-parameter (3PL) logistic model introduced by Birnbaum (1968) and used by TIMSS for multiple choice items. Letting X_{ig} represent the (0/1) response for individual i on item g, the IRF for the 3PL model is

$$P_g(X_{ig} = 1|\theta) = c_g + \frac{1 - c_g}{1 + \exp\left[-a_g\left(\theta - b_g\right)\right]}. \tag{1}$$

This function describes all 36 items administered to our sample and gives the probability of observing a correct response given ability θ and item parameters (a_g, b_g, c_g).

Figure 1 provides the intuition behind the 3PL model parameters. The *pseudo guessing parameter*, c_g, incorporates the fact that on multiple choice exams even the worst performers ($\theta \to -\infty$) will sometimes guess correctly. The *difficulty parameter*, b_g, measures the item's overall difficulty since the probability of answering correctly depends equally on ability and difficulty. The *discrimination parameter*, a_g, captures how quickly the likelihood of success changes with respect to ability. Intuitively, an item with a high discrimination parameter can distinguish between examinees with abilities just below and above b_g. Overall, this relatively flexible functional form has proved adept at fitting item response patterns.

To illustrate graphically how IRT links items and tests, Figure 2 plots the item response functions for two TIMSS items that map ability on the horizontal axis to the percentage correct on the test. A third curve plots the *test characteristic curve* for a test composed of these two items only. Since the item response functions are fully characterized by the published TIMSS items parameter and the structural assumption of a logistic function, it is easy to read the mean ability of a child by the percentage correct on the test. For instance, if item 19 is administered and 60 percent of children respond correctly, the mean ability is 425. By comparison, the same result on item 21 would suggest a higher mean ability level since the question is more difficult.

The key advantage of IRT in large testing situations is this ability to link tests, either in a cross-section (when different children are administered different test questions) or over time (when children are tested more than once). Formally, IRT equates competence levels by identifying off the set of common items across the tests and defining a reference population. Absent a reference population, the IRF given by (1) provides competence levels and item parameters that are identified up to an affine transformation—poor performance cannot be distinguished from a difficult test and a large variance in achievement cannot be distinguished from a highly discriminating test. Specifically, the transformations

$$\begin{aligned} \theta' &= \alpha\theta + k && (2)\\ b'_g &= \alpha b_g + k && (3)\\ a'_g &= \frac{a_g}{\alpha} && (4)\\ c'_g &= c_g && (5) \end{aligned}$$

will yield identical characteristic curves, so that $P_g(\theta'; a'_g, b'_g, c'_g) = P_g(\theta; a_g, b_g, c_g)$. However, *if item parameters are fixed*, the scale of θ—the mean and variance—is fixed as well. Thus by calibrating items using a defined reference group we can score the performance of all other children relative to that group, regardless of which items children actually complete. In our case, the reference group is given by the TIMSS knowledge scale. This scale fixes the item parameters such that the TIMSS 1995 sample of eighth grade children have mean 500 and standard deviation 100 (Yamamoto & Kulick 2000).

In our application, all students receive the same exam and all item parameters are fixed using TIMSS. In general, however, this need not be the case. Students can receive different exams and new items so long as each item can be linked to a common set of fixed items or a fixed reference population. For example, three two-item exams with item pairs (1,2), (2,3) and (3,4) can all be linked provided that one of the four items is fixed, even if each test is administered to a different population.[4]

[4] To see this, note that if we fix item 1—drawing it from the TIMSS item bank, say—we can estimate the

2.1 Estimating the Mean

Given a set of individuals who were administered the same test, the likelihood function of observing the $N \times G$ response matrix x is

$$L(\theta, a, b, c|x) = \prod_i \prod_g P_g(\theta_i; a_g, b_g, c_g)^{x_{ig}} \left[1 - P_g(\theta_i; a_g, b_g, c_g)\right]^{1-x_{ig}}, \tag{6}$$

where P_g is the 3PL model given by (1) and x_{ig} is the 0/1 response for individual i on item g. Because of convergence issues associated with joint maximum likelihood methods that iterate between solutions for item parameters and individual abilities, most researchers use marginal maximum likelihood (MML) to estimate the 3PL model. To estimate any unknown item parameters, this method integrates out the ability distribution $f(\theta)$ to get the marginal likelihood function. Bock & Aitken (1981) propose an efficient EM algorithm to perform the resulting maximization problem. In addition to the parameter estimates, this algorithm returns a summary measure of the ability distribution $f(\theta)$ such as a mean and variance or a quadrature approximation. To obtain individual ability estimates, one can maximize the full likelihood function (6) treating the item parameters as fixed. For our application, this is all that is required to produce MLEs since all parameters are known. The sample means—the average score in Orissa and Rajasthan—can be computed from the individual ability estimates or, potentially, the means obtained during the marginalization of the full distribution.

While maximum likelihood methods are usually perfectly adequate to estimate sample means, there are some exceptions. One significant problem is that MLE proficiency is undefined if children answer fewer items correctly than would be expected by chance. So long as one child has an undefined ability estimate, so too is the sample average. As a result, researchers commonly limit the proficiency scale to some finite range. We follow TIMSS and bound MLE scores between 5 and 995—in our sample, 91 of the 6000 tested children are bounded below by 5. A second, more technical concern relates to the methods used to maximize the likelihood function (6) for ability. Yen et al. (1991) find that this likelihood function is often multimodal even for tests

parameters for item 2 using the first exam. Given parameters for item 2, we can then estimate the parameter for item 3 using students who received the second test. These students need not have the same ability distribution as the first group because they can be compared directly using item 2. Using a similar argument we can link the third exam to the first two.

with up to 50 items, which is a potential pitfall for many numerical maximization algorithms commonly employed.

Bayesian methods avoid some of these problems by incorporating additional information through a prior. Leaving just enough notation to capture the basic idea, the Bayesian approach focuses on the *posterior* distribution,

$$p(\theta|x) \propto L(\theta|x)p(\theta), \tag{7}$$

which is proportional to the product of the likelihood and prior. The expected a posterior (EAP) estimate of ability is simply the mean of the posterior distribution for each individual θ_i. One advantage of EAP scores is that they are always well defined, even for poorly performing students; when the likelihood function provides no additional information, the posterior simply converges to the prior. Moreover, provided that the prior distribution is correctly specified, the mean of the EAP scores is an unbiased estimate of the sample mean and has a smaller mean squared error than the corresponding MLE based estimate.

2.2 Estimating the Variance and Quantiles

In addition to the average performance level in Rajasthan and Orissa, we are also interested in the shape of the full distribution. The primary difficulty here is that if the test is too short, too easy or too difficult, the individual errors become too large to ignore and the distribution of estimated individual proficiencies no longer converges to the population distribution (Yamamoto & Kulick 2000, Mislevy, Beaton, Kaplan & Sheehan 1992). To get a sense for whether this is an issue in the Indian case, Figure 3 plots the distribution of MLE abilities in a histogram (left axis) and the associated $\pm 1.96 * se$ confidence interval on the right axis.[5] For children below the mean, the precision of the ability estimate is very low. Simply put, for most Indian children, the test is too hard. In this situation, the mean of the sample will still generally approach the

[5] Item Response Theory provides the standard error for each score from the inverse Fisher information matrix after ML estimation of the IRT model. As the number of items grows large, this standard error summarizes the normal sampling distribution of the estimator. However, as the number of items shrinks, the sampling distribution becomes highly non-normal. In particular, our test is weakly informative for poorly performing students because we cannot distinguish between students scoring poorly and those score *very* poorly; we can reject that such students are high achievers. Consistent with how ML standard errors are calculated, Figure 3 does not capture this non-normal behavior and instead graphs $\pm 1.96 * se$.

population mean, but the same is not true for the estimated variance.

To see this, consider the variance of the MLE scores $\hat{\theta}$ and the EAP scores $\bar{\theta}$. The variance of the MLE scores includes both the variance of true scores θ and measurement error e. That is,

$$\operatorname{Var}(\hat{\theta}) = \operatorname{Var}(\theta) + \operatorname{Var}(e). \tag{8}$$

Defining the test reliability ratio as $\rho \equiv \operatorname{Var}(\theta)/\operatorname{Var}(\hat{\theta})$, we have

$$\operatorname{Var}(\hat{\theta}) = \frac{\operatorname{Var}(\theta)}{\rho} > \operatorname{Var}(\theta). \tag{9}$$

By comparison, the EAP scores are a weighted average of the MLE score and the population mean, $\bar{\theta} = \rho\hat{\theta} + (1-\rho)\mu$. The variance of the EAP scores is therefore

$$\operatorname{Var}(\bar{\theta}) = \operatorname{Var}\left(\rho\hat{\theta} + (1-\rho)\mu\right) = \rho^2 \operatorname{Var}(\hat{\theta}) = \rho \operatorname{Var}(\theta) < \operatorname{Var}(\theta). \tag{10}$$

The true variance, $\operatorname{Var}(\theta)$, is bounded above by the MLE score variance and below by the EAP score variance. It should be clear that this argument extends to any percentile moments such as the top and bottom quintile. Unfortunately, the error structure in IRT is complicated and closed-form corrections are not readily available.

One simple way to address this issue is to bound the distribution estimates using MLE and EAP scores. Where these estimates are similar, no further work may be required—convenient because both MLE and EAP scores are readily available from standard reports in test analysis programs such as BILOG-MG. Unfortunately, in parts of the distribution where the test is only weakly informative the bounds may be quite large; in our application, this turns out to be true for estimates of lower quantiles.[6]

A more satisfactory solution, and the one followed by TIMSS, is to draw "plausible values" from the posterior distribution of each student's ability estimate and then use these draws to approximate the true achievement distribution (Mislevy 1991, Mislevy, Beaton, Kaplan & Sheehan 1992, Mislevy, Johnson & Muraki 1992, Yamamoto & Kulick 2000). Staying with our simplified

[6] Another approach would be to use the information obtained during the integration process of the marginal maximum likelihood procedure. While this would give unbiased estimates of certain moments it depends crucially on the integration strategy used and is generally rather cumbersome.

posterior notation, we draw five plausible values for each child

$$\tilde{\theta}_{i,k} \sim p(\theta_i|x) \qquad\qquad (k = 1, ..., 5) \tag{11}$$

and then estimate sample moment of interest as

$$\hat{s} = \frac{1}{5}\sum_{k=1}^{5} s(\tilde{\theta}_k) \tag{12}$$

where $s(\tilde{\theta}_k)$ may be the variance, 90th percentile, etc, of the N element vector of plausible values $\tilde{\theta}_k$.

Unfortunately, no publicly available software can draw plausible values for the model we estimate, making it difficult for other researchers to replicate the TIMSS methodology precisely. We use the Markov Chain Monte Carlo (MCMC) algorithm proposed by Patz & Junker (1999*a*,*b*) to compute the EAP scores and plausible values. This technique differs from the computational approach used by TIMSS but is highly flexible and relatively straightforward to implement. We provide a fuller explanation of our estimation strategy in Appendix A.

To see whether these concerns are of practical importance, Figure 4 shows the estimated MLE, EAP and PV distributions of ability. The MLE and EAP estimates represent the upper and lower bounds of the variance of the distribution, with the PV estimates somewhere in between. There is a considerable divergence in the shape of the distribution to the left of the mean, while at higher levels of ability, the three distributions look roughly similar. Because individual scores are only weakly informative for poor performers, the MLE and EAP estimates diverge and plausible value methodology is crucial to estimating percentile cutoffs at the bottom of the distribution.

3 Data

We use data collected in 2005 as part of a larger World Bank study designed and led by Kin Bing Wu, and conducted by the Social and Rural Research Institute (SRI) unit of IMRB International.[7] The study includes detailed surveys of principals, teachers, and students in 288

[7] For further details on the survey design and an analysis of the correlates of performance, see Wu et al. (2006, 2007).

schools in the Indian states of Rajasthan (in the West) and Orissa (in the East). The sample was designed by first selecting districts using population proportional to size (PPS) sampling, where, in the absence of data on school-by-school enrollment, the PPS methodology was applied to the population of schools across districts. Following the selection of districts, the total number of schools to be selected in each district was arrived at and schools were surveyed in both urban and rural areas, as well as across institutional affiliation; that is, government schools, private unaided and private aided schools were all included in the sample.

There are three important features of the data. First, only children enrolled in school were tested. According to the World Development Indicators, gross enrollment in India in lower secondary schools is only 53 percent, so this leaves out half the children of the relevant age-group. Consequently, there are intractable issues of trying to generalize the comparison of tested children to *all* children across countries. For instance, the gross enrollment in secondary schools in three other countries close to Orissa and Rajasthan in the world rankings varies dramatically, from 44 percent in Ghana to 75 percent in Botswana and 90 percent in South Africa. To the extent that marginal performers are less likely to be enrolled, comparisons of India with South Africa would therefore favor India; alternatively comparisons between India and Ghana favor the latter. Although problematic for the mean, the lack of information on non-enrolled children may not be as problematic for different percentiles—it may be plausible to assume, for instance, that the 50 percent of children not tested are likely to join the group that performs "poorly", in a sense to be made precise below.

Second, although all attempts were made to ensure that no type of school or location was left out of the sampling procedure, it has been difficult to accurately weight the data given paucity of data on enrollments in private unaided and aided schools at the district level. This is a general problem that any testing exercise has to address and it calls for an urgent compilation of a universal dataset that can be used for sampling in the future.

Third, the data are from two states only, and therefore generalizations to all of India may be misleading—Rajasthan and Orissa are both poorer states with larger tribal populations. Learning outcomes though may be different from those suggested by income rankings. The results from a countrywide testing exercise in rural areas (Pratham 2006) gives us some sense of where these states lie in the Indian distribution. Among children tested in Grade 8 countrywide (rural areas only), 82.4 percent could read a story, 75.2 percent could divide and 95.5 percent could write. The average of Orissa (83.98 percent (read), 71 percent (divide) and 95.3 percent (write)) and Rajasthan (92.9 percent can read, 92.4 percent can divide and 98.5 percent can write) is surprisingly not far off the Indian average—if anything, these results suggest that children in these two states may be scoring higher than the rest of the country. However, lots of caution is still warranted—particularly since Orissa performs better than Rajasthan in the tests we use while Pratham finds the opposite.

In the selected schools, students in ninth grade were administered a 36-item test where all items were selected from the list of publicly released items published by the TIMSS. The test sought to cover the content domains tested under the TIMSS with 11 items on Algebra, 5 on Data Representation, Analysis and Probability, 9 on Fractions and Number Sense, 7 on Geometry and 4 on Measurement. The performance expectation across these content domains also varied and ranged from "Communicating and Reasoning" to "Using Complex Procedures" (Table 1). The items selected were neither too difficult nor too hard in the TIMSS calibration, ranging from -1.07 (a student 1 standard deviation below the mean would have a 50 percent chance of answering this question correctly, absent guessing) to 1.244; the items were also uniformly distributed across this difficulty range.

4 Results

4.1 International Benchmarking

There are two views that currently dominate thinking about educational policy in India. One view—active proponents of which include prominent NGOs—is that Bharat is drowning. Average learning levels are so low that the typical child will leave primary school without knowing how to read or perform elementary mathematical operations. A second view—often expressed by those in the government and in the media—is that India is shining. This group points to India's increasing global presence, the large number of Indian professionals in high paying jobs, and the dramatic growth of its service industry, particularly in information technology. As it turns out, both views contain an element of truth, and both views can be justified by presenting different pieces of the same data.

Mirroring the view that Bharat is drowning, absolute achievement, as measured by the percent correct score, is low compared to curricular standards. A significant fraction of children have not mastered the content categories expected for their grade (Table 1). By ninth grade, only 11 percent of children in Rajasthan and 17 percent in Orissa can correctly choose the smallest value from the set 0.625, 0.25, 0.375, 0.5, and 0.125 (Table 1, Q2). The question "A runner ran 3000m in exactly 8 minutes. What was his average speed in meters per second?" (Q7) stumps all but 23 percent of children in Rajasthan and 32 percent of children in Orissa. A simple test of division and fractions—"if 4 times a number is 48 what is 1/3 of the number?" (Q10)—is too difficult for 65 percent of children in Rajasthan and 64 percent of children in Orissa.

While the item-by-item comparison suggests that Indian children are performing significantly below the international average, interpreting the magnitude of this effect is difficult because it depends on a test-specific metric. As discussed, the percentage correct score is a function of latent achievement differences—our true parameter of interest—and the discriminating power of the test, and thus inseparable from the specific test design.

Figure 5 uses the linking methodology proposed previously to present cross-country comparisons on the TIMSS achievement scale.[8] Based on the average score, Rajasthan and Orissa rank below 46 (42) of the 51 countries tested with a score of 382 and 404. This ranking straddles Bahrain, Chile and Morocco and is boxed in by Egypt above and the Philippines below. Averaging across the entire tested sample, India scores 392—below 43 of 51 countries. This compares to the international average of 487 in 1999 and 467 in 2003. Seen in standard deviations of all children tested, the two Indian states are 0.7 student standard-deviations or 1 country standard-deviation below the TIMSS mean. Relative to the OECD mean, the tested Indian children are 3.1 (OECD) standard deviations below. That is, if we rank all the OECD countries, India would lie below the 1st percentile in the distribution of OECD country scores.

The true picture may be worse. Since the tests included only enrolled children, the comparisons favor India to the extent that enrollment is lower relative to other countries. In both

[8] We follow the TIMSS methodology as closely as possible and compute sample averages using the EAP scores, which is, in this case, simply more efficient than using plausible values. The MLE scores, which are estimated using BILOG-MG rather than our custom MCMC routines, yield somewhat lower estimates of the average: 374 and 386. The discrepancy between the EAP and MLE averages is likely due to students scoring in an area where the likelihood function is virtually flat or undefined. In this situation, regularity and stability become a major concern with MLE.

Botswana (75 percent) and South Africa (90 percent) gross enrollment in secondary schools is higher. It is likely that a representative sample of children (enrolled and unenrolled) would place India below additional countries.

That the average child is performing poorly masks the considerable variation in the distribution. At the bottom, children score extremely poorly. There is no evidence that the distribution is more compressed at the bottom than for other low-performing countries. In fact, only three countries—Saudi Arabia, Ghana, and South Africa—score worse than Rajasthan or Orissa if ranked by the 5th percentile cutoff score (Figure 6). When the education system fails, it fails completely.

4.2 Inequality in the Learning Distribution

Following Micklewright & Schnepf (2006), we report a simple statistic measuring test-score dispersion—the difference between 5th and 95th percentiles of the test score distribution. Figure 7 shows the significant educational inequality in the Indian learning distribution. In both the Indian states, the 5-95 percentile spread is greater than 300, and just below the most unequal country in the TIMSS sample—South Africa.

TIMSS 2003 also presents achievement benchmarks based on an intensive effort to anchor performance to objective criteria. Table 2, drawn from TIMSS 2003 (Exhibit 2.1), describes the low (400), intermediate (475), high (550), and advanced (625) international benchmarks; Table 3 shows the results. In Rajasthan and Orissa, 1 percent of children pass the advanced benchmark. This actually is above many other poor performing countries. At the same time, only 42 percent in Rajasthan and 50 percent in Orissa pass the lowest benchmark. Put another way, only 40 to 50 percent of Rajasthan and Orissa's enrolled ninth graders have "some basic mathematical knowledge"—the description of the low international benchmark.

A second useful exercise that demonstrates the vast differences between tested children is to rank Table 3 by those who reach each of the different international benchmarks. Ranked by the low international benchmark, Rajasthan is 8th from the bottom and Orissa 9th; ranked by the intermediate benchmark, they are now 9th and 14th from the bottom respectively; ranked by the high international benchmark they are now 11th and 16th from the bottom. The advanced international benchmarks put both states at the respectable positions of 12th and 18th, although the precise ranking is difficult to obtain given rounding.

To the extent that these two states represent India, the combination of a wide achievement distribution and immense population explains why perceptions of India can vary so dramatically. In Table 4, we use population age-cohort estimates and enrollment rates to estimate the *number* of 14-year olds in each country who pass the international benchmarks set by TIMSS. The results are striking. If one percent of Indian children reach the advanced international benchmark—the average suggested by Rajasthan and Orissa—the total cohort size ranks 5th out of all the countries tested. Only Japan, the United States, South Korea, and Taiwan have more students passing the top benchmark. For every ten children who pass the advanced benchmark in the United states, there are four children who pass the benchmark in India. Indeed, the 101 thousand Indian children who pass the advanced benchmark exceeds the total number of children who pass in bottom 32 countries combined. If India were added to the TIMSS sample, one out of every fourteen children who pass the advanced benchmark would be Indian.

The view from the top—that Indian's form a substantial fraction of top performers worldwide—contrasts sharply with the view from the bottom. The sheer magnitude of India's youth population and poor average performance means that over 17 million Indian 14-year olds are either not enrolled or fail the low international benchmark. This number is 22 times the number in the United States, 217 times the number in South Korea, and 726 times the number Japan. Indeed, there are more Indian's either not enrolled or who fail the low benchmark than in all the other TIMSS countries combined.

4.3 Variance Decomposition

The striking disparity between top- and bottom-achievers hints that children receive different educational inputs, both based on the state in which they live and the characteristics of their families and schools. While it is impossible to draw definitive causal conclusions using simple correlations or variance decompositions, the patterns that emerge from even a basic analysis are broadly consistent with a view of an education system rife with inequality but rich in potential. In a hopeful sign, the form inequality takes suggests that public policy plays a role. The impact of household attributes—educational inputs that the government has little power to control—appears mitigated by the institutional structure of states and schools.

We present a heuristic approach towards examining the source of achievement in Figure 8. Here, we first regress test scores on district dummies and then plot the residuals—this is a measure of how much of the variation is accounted for by districts. We then add in child and household characteristics—age, gender, caste, parental literacy, and wealth—and plot the residuals again; finally we repeat the exercise including school dummies. To the extent that districts, households, or schools explain a large portion of the variation in the test score data, we expect that residual plot to be more "concentrated" once the appropriate dummies are accounted for. So, if districts matter a lot, we expect the residual plot from a regression of test scores on district dummies to be "tighter" than the distribution of all test scores.

As Figure 8 shows, schools seem to matter most. Progressively adding district effects and family characteristics compresses the distribution slightly. Only when we add school fixed effects is the collapse noticeable; the gaps between schools accounts for more than the gaps between children from different household characteristics.

Table 5 confirms this result more formally using a simple regression based variance decomposition. Here, we first regress achievement on district dummies. The R^2 from this regression gives a measure of the variance explained by districts alone. Examining the change in R^2 after adding household controls gives the fraction of achievement variation explained by observable characteristics above and beyond the district effect. While indicative of households' contribution to learning, we cannot claim households causally explain this fraction of variance since children sort into schools. If this occurs, observable household characteristics may explain achievement simply because schools determine learning and children sort. Proceeding onward, we add school dummy variables and report the increase in R^2. This gives some sense for the importance of schools, but again we cannot make definitive causal statements. A significant increase in variance explained at this stage implies either that schools matter or that children sort on unobservable characteristics. After accounting districts, observables, and schools, the remaining variation is idiosyncratic. As Figure 8 shows, measurement error, which cannot be decomposed by definition, forms a significant portion of this idiosyncratic variation.

Table 5 shows the results of this exercise. In Orissa (Rajasthan), schools explain an additional 32 percent (41 percent) of the test score variation above districts and observable household characteristics. This is twice the amount of variation explained by districts and household characteristics in Orissa and five times the variation explained by those attributes in Rajasthan. Even if half of this effect is due to selection on unobservables, schools remain important. For comparison, the maximum variation possibly attributable to school specific factors in OECD countries is 14 percent—less than half the value for India (Pritchett 2004). If we were to remove the variation due to measurement error and renormalize our decomposition to sum to one, the schools' role would appear even more significant.

5 Robustness Checks

Some caveats are in order. TIMSS uses a complex test design where children are given a subset of items in a specific format. Our results are based on a test that includes 36 TIMSS questions, but the test-design is clearly different. The educational testing literature has many examples of design effects, where test scores are shown to change depending on the design of the test. By presenting results using IRT equating methods, we are essentially ignoring this rich literature.

One robustness check used in the item response literature compares the actual responses of children, averaged across ability groups with that predicted on the basis of item parameters. In our particular case, these tests of "item fit" reveal the extent to which the shape of the item response function *predicted from the TIMSS item parameters* corresponds to the actual responses of examinees. Figure A1 shows the predicted and actual responses for all 36 items.

For the majority of items, both the 3PL model and the item parameters closely predicted how children would perform. In a few instances, however, the fit could be improved. As an example, item 33 is a poorly-fitted item where high ability Indian children seem to struggle more than their international peers. While these few items are unlikely to introduce significant bias, future researchers should carefully select items during the pilot phase to minimize deviations from the expected response patterns.

Further, a factor model of item responses generated the first eigenvalue (3.9) 9 times greater than the second (0.4), easily satisfying Drasgow & Lissak's (1983) rule-of-thumb for assessing the unidimensionality assumption. Nevertheless, we could not conduct formal tests of Differential Item Functioning (DIF) given that we do not have access to item-by-item responses for other TIMSS examinations (and these are typically not available in the public domain). Mullis &

Martin (2000), however, conduct the required analysis for the TIMSS 1999 sample and there is little reason to suspect the results would not extend to India.

The methods and results discussed here should not be taken as advocacy for dispensing with TIMSS altogether and using their publicly released items to place tested children on international distributions. TIMSS provides a level of analysis and robustness checking that independent researchers cannot easily replicate. We view the methods presented here more as a bridge between current practices and TIMSS-like comparability rather than an alternative. Even in this case, a larger pilot that compares TIMSS results with those obtained by the methods suggested here would yield important information on the biases inherent in our equating methods.

6 Conclusion

The educational administration in India has often shaken off the bad news emerging from the primary educational sector on the grounds that the Indian system is based on the rigors of selection. A gruelling primary schooling would weed out all but the best performers, who would then graduate onwards to secondary schools and receive a higher quality education. One response to the poor testing results from the primary level has in fact been to point to India's position in the global economy and the comparable performance of its top firms and professionals to their international counterparts. In essence, if the schooling system is so poor, how is it that India has all these top global performers?

But this misses the point. Both positions are sustained by the data. Children from these two states clearly fail any potential Millennium Learning Goal. If results are similar for the rest of the country, over 17 million 14 year-olds, around 80 percent of the population, are either not enrolled or cannot pass the lowest international benchmark. But India's massive population and wide variance in achievement also ensure that Indians are amply represented in the worldwide cohort of top performers. One out of fourteen children who pass the advanced benchmark in the TIMSS sample are Indian, a ratio only four other countries can match. For every ten children in the United States who pass the advanced benchmark—and only 7 percent do—there are four who pass it in India.

How this situation plays out over the next decade has much to do with how production technologies evolve in the labor market. If Indian firms manage to adopt "Ford Model-T" technologies that require a handful of highly skilled and educated workers to match with a large number of unskilled workers, India shining can act as a "rising tide that lifts all boats." But if Indian firms adopt "McKinsey" technologies that require skilled workers and unskilled workers to match among themselves (as the IT consulting firms require, but not necessarily call-centers) it is likely that the country will be characterized by increasing inequalities; an enclave of a few privileged and self-perpetuating rich surrounded by a majority poor.

There is some hope in the variance decompositions and associations that inequalities in the educational system can be addressed through government policies. A consistent finding across OECD countries is the low explanatory power of schools in explaining the variation in test scores compared to households. This is problematic for policy, since it is easier to change behavior among teachers and to improve schools, than it is to do the same among parents. That a large fraction of the variation in achievement arises from differences across schools suggests that there are school-level variables, manipulable by policy, that could result in positive impacts. What these might be, and where to go from here, should form the basis of future research and evaluations.

More generally, the methods proposed in this paper highlight the potential benefits of linking scores to the worldwide achievement distribution. While such efforts cannot replace the important work undertaken by TIMSS, they represent a clear improvement over the collection of ad-hoc exams employed by most researchers, and require little additional work. India is hardly alone in its absence from the TIMSS rankings, and many countries could benefit from an analysis similar to ours. Over time, through such efforts, independent researchers may help make tracking a Millennium Learning Goal a reality.

A Item Response Theory

A.1 Estimating MLE Scores

Linking our test form to the TIMSS knowledge score distribution requires a underlying model of the response process. In our case, all 36 items presented can be described by the 3PL model given (1). Letting $x_{ig} \in \{0,1\}$ denote the response for individual i on item g and X be the full data matrix, the likelihood of observing X given a vector of associated abilities, θ, is

$$\begin{aligned} P(X|\theta) &= \prod_{i}^{N}\prod_{g}^{G} P_g(x_{ig}|\theta_i) & (13) \\ &= \prod_{i}^{N}\prod_{g}^{G} P_g(\theta_i)^{x_{ig}}[1-P_g(\theta_i)]^{1-x_{ig}}, & (14) \end{aligned}$$

where the product form arises from assuming independence across items and individuals. Unlike most IRT models we have suppressed the notation of the item parameters to highlight the fact that they are fixed. In many cases there may be a mix of fixed anchor items and new uncalibrated items, but we do not face that situation here.

With fixed parameters it is relatively trivial to maximize the likelihood function associated with each individual using Newton-Raphson or some other numerical procedure; each first order condition is independent of the others so we do not face a curse of dimensionality. But some difficulties remain. In particular, the 3PL model's guessing parameter makes MLEs undefined for those scoring below the guessing rate. These flat parts of the likelihood function can make numerical estimates unstable. Yen et al. (1991) also find that some response vectors can produce likelihood functions with multiple modes even for tests of a reasonable length (such as 36 items). These modes can trap derivative based maximization algorithms at local rather than global peaks. To study these issue, we computed ML estimates using both a Newton-Raphson algorithm and BILOG-MG. While the estimates agreed perfectly for most individuals, there appeared to be some instability, particularly near the bottom of the distribution where our test is only weakly informative and where students often score below the guessing rate. Given these differences we choose to report only BILOG based ML estimates.

A.2 Estimating EAP Scores and Plausible Values by Markov Chain Monte Carlo

Both EAP and plausible values are based on the posterior distribution of individuals' ability. In Section 2 we introduced the basics of the Bayesian approach using simplified notation. To be more precise, we now change the setup slightly and introduce notation for manifest predictors of the score. Letting Y denote the matrix of predictors such as state, gender, age, wealth, parental literacy and school type, we follow TIMSS and assume that covariates are linked to ability using a simple linear model

$$\theta = Y\beta + \epsilon, \tag{15}$$

where $\epsilon_i \sim N(0, \sigma^2)$. Given this model, we can express the joint posterior distribution for all parameters as

$$\begin{aligned}
P(\theta, \beta, \sigma | X, Y) &\propto P(X|\theta, \beta, \sigma, Y) P(\theta, \beta, \sigma | Y) && (16)\\
&= P(X|\theta) P(\theta, \beta, \sigma | Y) && (17)\\
&= P(X|\theta) P(\theta | \beta, \sigma, Y) P(\beta, \sigma | Y) && (18)\\
&= P(X|\theta) P(\theta | \beta, \sigma, Y) P(\beta) P(\sigma) && (19)\\
&= \left[\prod_i \left(\prod_j P_j(x_{ij}|\theta_i) \right) P(\theta_i | \beta, \sigma, Y_i) \right] P(\beta) P(\sigma) && (20)
\end{aligned}$$

where (16) follows from Bayes Rule, (17) follows from unidimensionality, (18) follows from the multiplication rule, (19) follows from independence of β, σ^2 and Y, and (20) follows from the independence across individuals and items. Our parameters of interest—the EAP and plausible value estimates of ability—are the expected value of the posterior θ_i or simply independent draws from this distribution. One can therefore think of plausible values as an empirical approximation of the posterior.

The computational problem becomes how to draw from this posterior distribution. Patz & Junker (1999*a*,*b*) illustrate how Markov Chain Monte Carlo (MCMC) techniques, particularly so-called Metropolis-Hastings within Gibbs, can be used to draw from the posterior distribution even in very complicated IRT settings. The basic idea of MCMC is to simulate observations from a Markov chain whose stationary distribution is the joint posterior distribution of interest. There are many strategies for constructing a chain with this property. In the IRT context, MH-within-Gibbs achieves the objective in a relatively straightforward manner.

The basic motivation behind "Gibbs samplers" is to reduce the simulation problem to lower dimensional, perhaps univariate, space. In our case, we are interested in the distribution of $N + K + 1$ random variables, $\pi = \theta_1, \ldots, \theta_N, \beta_1, \ldots, \beta_K, \sigma | X, Y$. Gibbs sampling constructs a Markov chain $M_t = (\theta_1^{(t)}, \ldots, \theta_N^{(t)}, \beta_1, \ldots, \beta_K^{(t)}, \sigma^{(t)})$ by sampling from the *full conditionals* as follows:

- $\theta_1^{(t+1)} \sim p(\theta_1 | \theta_2^{(t)}, ..., \theta_N^{(t)}, \beta_1, ..., \beta_K^{(t)}, \sigma^{(t)}, X, Y)$
- $\theta_2^{(t+1)} \sim p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_N^{(t)}, \beta_1, ..., \beta_K^{(t)}, \sigma^{(t)}, X, Y)$
- $\vdots$

It can be shown this chain converges to a stationary distribution π (e.g. Casella & George 1992, Tierney 1994). In the IRT context the full conditionals simplify considerably because of independence between individuals. That is, transition probabilities for each type of $N + K + 1$ parameters is given by

$$p(\theta_i | \beta, \sigma, X_i, Y_i) = \frac{\left(\prod_j P(x_{ij} | \theta_i)\right) P(\theta_i | \beta, \sigma, Y_i) P(\beta) P(\sigma)}{\int \left(\prod_j P_j(x_{ij} | \theta_i)\right) P(\theta_i | \beta, \sigma, Y_i) P(\beta) P(\sigma^2) d\theta_i} \tag{21}$$

$$p(\beta_k | \theta, \beta_{-k}, \sigma, X_i, Y_i) = \frac{\left[\prod_i \left(\prod_j P(x_{ij} | \theta_i)\right) P(\theta_i | \beta, \sigma, Y_i)\right] P(\beta) P(\sigma)}{\int \left(\prod_j P_j(x_{ij} | \theta_i)\right) P(\theta_i | \beta, \sigma, Y_i) P(\beta) P(\sigma^2) d\beta_k} \tag{22}$$

$$p(\sigma | \theta, \beta, X_i, Y_i) = \frac{\left[\prod_i \left(\prod_j P(x_{ij} | \theta_i)\right) P(\theta_i | \beta, \sigma, Y_i)\right] P(\beta) P(\sigma)}{\int \left(\prod_j P_j(x_{ij} | \theta_i)\right) P(\theta_i | \beta, \sigma, Y_i) P(\beta) P(\sigma) d\sigma} \tag{23}$$

If sampling from these full conditional distributions is easy, Gibbs sampling provides a means to generate a sample from the posterior of each parameter.

In practice, computing the normalizing constant in the denominator of each conditional may be difficult—e.g. a closed form solution may not exist. The MH-within-Gibbs algorithm avoids this complication by inserting a Metropolis step when sampling from the full conditionals. Chib &

Greenberg (1995) provide an excellent pedagogic introduction to Metropolis-Hastings algorithms. A representative example of the algorithm for parameter θ_i is:

1. Simulate

$$\tilde{\theta}_i \sim \theta_i^{(t)} + \nu_i \quad \nu_i \sim N(0, s_i) \tag{24}$$

2. Accept the proposed value as follows:

$$\theta_i^{(t+1)} = \begin{cases} \theta_i^{(t)} & \text{with probability } 1-\alpha \\ \tilde{\theta}_i & \text{with probability } \alpha \end{cases} \tag{25}$$

where

$$\alpha = \min\left\{1, \left(\frac{p(\tilde{\theta}_i|\beta^{(t)}, \sigma^{(t)}, X_i, Y_i)}{p(\theta_i^{(t)}|\beta^{(t)}, \sigma^{(t)}, X_i, Y_i)}\right)\right\} \tag{26}$$

By using a symmetric proposal distribution $N(0, s_i)$ the normal MH criterion α does not include the the proposal distribution. Moreover, note that by substituting (21) into (26) we are left with an algorithm that includes only known functions since the denominator cancels. We can therefore easily compute α and simulate a Markov chain that converges to the posterior of interest. The MH steps for the regression parameters β_k and σ are completely analogous. For a more comprehensive description of MCMC methods applied to IRT problems see Patz & Junker (1999*a*,*b*).

To compute the EAP and plausible values estimates we ran a chain of 4,000 observations, discarding the first 2,000 as a burn-in period. As part of the linear model, we included private school attendance, age, age-squared, family size, family size squared, gender, father literacy, mother literacy, wealth category, caste, state, school facilities category, an intercept and a missing data dummy as explanatory variables. Including these manifest predictors makes our estimates more precise and is required for subsequent analysis using plausible values to be valid (Mislevy, Beaton, Kaplan & Sheehan 1992). We assumed flat priors for the β and σ parameters making the EAP estimates analogous to empirical Bayes, although this assumption has little effect since the data dominates the prior for these parameters. To ensure convergence, we experimented with the proposal distribution variances until the acceptance rates average around 44 percent with no significant outliers. Visually checking the chain graphs and running multiple chains and comparing the results confirmed that the chains rapidly converged after several hundred observations and autocorrelations were modest. Finally, we averaged the last 2,000 observations to compute the the EAP estimate. Even with this relatively modest chain length, the Monte Carlo error was tiny compared to the variance associated with each score. We also took five evenly spaced draws from the posterior as plausible values.

References

Aggarwal, Y. (2000), *Primary Education in Delhi. How much do the Children learn?*, NIEPA, New Delhi.

Bashir, S. (1994), 'Achievement Performance at the Primary Level in Public and Private Schools of Tamil Nadu', *Indian Education Review* **29**(3-4), 1–26.

Bedard, K. & Ferrall, C. (2003), 'Wage and test score dispersion: some international evidence', *Economics of Education Review* **22**(1), 31–43.

Birnbaum, A. (1968), Some Latent Trait Models and Their Use in Inferring an Examinee's Ability, *in* F. M. Lord & M. R. Novick, eds, 'Statistical Theories of Mental Test Scores', Addison-Wesley Publishing Company.

Blau, F. D. & Kahn, L. M. (2005), 'Do Cognitive Test Scores Explain Higher US Wage Inequality?', *The Review of Economics and Statistics* **87**(1), 184–193.

Bock, R. & Aitken, M. (1981), 'Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm', *Psychometrika* **46**, 443–459.

Brown, G. & Micklewright, J. (2004), 'Using International Surveys of Achievement and Literacy: A View from the Outside', *UNESCO Institute for Statistics Working Paper* .

Casella, G. & George, E. (1992), 'Explaining the Gibbs Sampler', *The American Statistician* **46**(3), 167–174.

Chib, S. & Greenberg, E. (1995), 'Understanding the Metropolis-Hastings Algorithm', *The American Statistician* **49**(4), 327–335.

Chinapah, V., H'ddigui, E. M., Kanjee, A., Falayajo, W., Fomba, C. O., Hamissou, O., Rafalimanana, A. & Byomugisha, A. (2000), *With Africa for Africa. Towards Quality Education for All*, Human Sciences Research Council, Pretoria, South Africa.

Debroy, B. & Bhandari, L. (2007), 'Exclusive growth – inclusive inequality', *Center for Policy Research Working Paper* .

Devroye, D. & Freeman, R. (2001), ‘Does Inequality in Skills Explain Inequality in Earnings Across Advanced Countries?’.

Drasgow, F. & Lissak, R. (1983), ‘Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses’, *Journal of Applied Psychology* **68**(3), 363–73.

Filmer, D., Hasan, A. & Pritchett, L. (2006), ‘A Millennium Learning Goal: Measuring Real Progress in Education’, *CGD Working Paper 97* .

Govinda, R. & Varghese, N. (1993), *Quality of Primary Schooling in India: A Case Study of Madhya Pradesh*, International Institute for Educational Planning, UNESCO; NIEPA, National Institute of Educational Planning and Administration.

Goyal, S. (2007), Learning achievements in india: A study of primary education in rajasthan, Technical report, Human Development Unit, South Asia Region. The World Bank.

Hasan, A. (1995), *Baseline Survey of Learning Achievement in Primary Grades in Bihar*, AN Sinha Institute of Social Studies, Patna.

Jalan, J. & Murgai, R. (2007), “‘intergenerational mobility in education in india’, *Processed. Delhi: the World Bank* .

Kijima, Y. (2006), ‘Why did wage inequality increase? Evidence from urban India 1983–99’, *Journal of Development Economics* **81**(1), 97–117.

Micklewright, J. & Schnepf, S. V. (2006), ‘Inequality of learning in industrialized countries’, *IZA Discussion Paper No. 2517* .

Mislevy, R. (1991), ‘Randomization-based inference about latent variables from complex samples’, *Psychometrika* **56**(2), 177–196.

Mislevy, R., Beaton, A., Kaplan, B. & Sheehan, K. (1992), ‘Estimating Population Characteristics from Sparse Matrix Samples of Item Responses’, *Journal of Educational Measurement* **29**(2), 133–161.

Mislevy, R., Johnson, E. & Muraki, E. (1992), 'Scaling Procedures in NAEP', *Journal of Educational Statistics* **17**(2), 131–154.

Mullis, I. V. & Martin, M. O. (2000), Item Analysis and Review, *in* M. O. Martin, K. D. Gregory & S. E. Stemler, eds, 'TIMSS 1999 Technical Report', International Study Center Boston College, Chestnut Hill, Massachusetts, pp. 225–234.

Nickell, S. (2004), 'Poverty and Worklessness in Britain', *Economic Journal* **114**(494), C1–C25.

Patz, R. & Junker, B. (1999*a*), 'A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models', *Journal of Educational and Behavioral Statistics* **24**(2), 146.

Patz, R. & Junker, B. (1999*b*), 'Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses', *Journal of Educational and Behavioral Statistics* **24**(4), 342.

Pratham (2006), *Annual Status of Education Report*, Pratham, Mumbai.

Pritchett, L. (2004), 'Towards a New Consensus for Addressing the Global Challenge of the Lack of Education', *Copenhagen Consensus Challenge Paper* .

Shukla, S., Garg, V., Jain, V., Rajput, S. & Arora, O. (1994), *Attainment of Primary School Children in India*, National Council of Educational Research and Training, New Delhi.

Strauss, J. & Burger, M. (2000), *Monitoring Learning Achievement Project*, Pretoria: Department of Education.

Tierney, L. (1994), 'Markov Chains for Exploring Posterior Distributions', *The Annals of Statistics* **22**(4), 1701–1728.

Wu, K. B., Goldschmidt, P., Boscardin, C. K. & Azam, M. (2007), Girls in india: Poverty, location and social disparities, *in* M. A. Lewis & M. E. Lockheed, eds, 'Exclusion, Gender and Education: Case Studies from the Developing World.', Center For Global Development, Washington D.C.

Wu, K. B., Goldschmidt, P., Boscardin, C. K. & Sankar, D. (2006), Student achievement in mathematics and its determinants in rajasthan and orissa, *in* 'Report on the Survey of Public and Private Secondary and Senior Secondary Schools', The World Bank. Processed.

Yamamoto, K. & Kulick, E. (2000), Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales, *in* M. O. Martin, K. D. Gregory & S. E. Stemler, eds, 'TIMSS 1999 Technical Report', International Study Center Boston College, Chestnut Hill, Massachusetts, pp. 237–264.

Yen, W., Burket, G. & Sykes, R. (1991), 'Nonunique solutions to the likelihood equation for the three-parameter logistic model', *Psychometrika* **56**(1), 39–54.

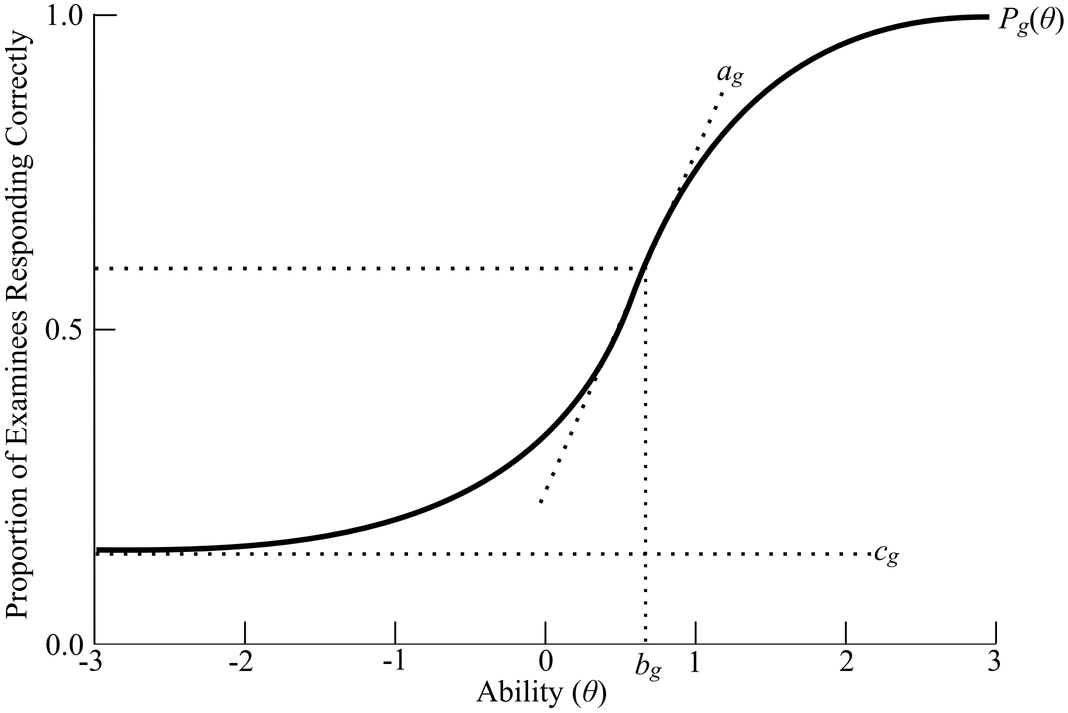
Table 1. Item Comparison of 2005 Assessment of 9th Grade Mathematics in Rajasthan & Orissa with TIMSS 1999 Assessment of 8th Grade Mathematics

Q No.	Content Category	Performance Expectation	Rajasthan All Students	Orissa All Students	International Average
No. 1	Data Representation, Analysis and Probability	Using Complex Procedures	33	32	60
No. 2	Fractions and Number Sense	Knowing	11	17	46
No. 3	Geometry	Using Complex Procedures	26	31	59
No. 4	Algebra	Knowing	48	47	65
No. 5	Geometry	Investigating and Solving Problems	39	48	62
No. 6	Algebra	Knowing	32	30	50
No. 7	Fractions and Number Sense	Investigating and Solving Problems	23	32	33
No. 8	Data Representation, Analysis and Probability	Knowing	43	24	57
No. 9	Measurement	Knowing	32	20	49
No. 10	Algebra	Investigating and Solving Problems	35	36	47
No. 11	Fractions and Number Sense	Knowing	30	21	50
No. 12	Data Representation, Analysis and Probability	Using Complex Procedures	54	46	64
No. 13	Algebra	Knowing	24	40	49
No. 14	Measurement	Investigating and Solving Problems	29	36	42
No. 15	Geometry	Knowing	38	48	54
No. 16	Fractions and Number Sense	Using Routine Procedures	16	26	39
No. 17	Geometry	Using Routine Procedures	36	36	58
No. 18	Algebra	Using Routine Procedures	38	51	65

No. 19	Fractions and Number Sense	Using Complex Procedures	55	54	75
No. 20	Data Representation, Analysis and Probability	Using Complex Procedures	43	39	58
No. 21	Algebra	Communicating and Reasoning	28	39	45
No. 22	Algebra	Using Routine Procedures	32	36	33
No. 23	Geometry	Investigating and Solving Problems	23	31	40
No. 24	Fractions and Number Sense	Using Routine Procedures	40	35	52
No. 25	Fractions and Number Sense	Knowing	47	49	61
No. 26	Measurement	Knowing	51	49	60
No. 27	Fractions and Number Sense	Investigating and Solving Problems	32	37	44
No. 28	Measurement	Investigating and Solving Problems	19	31	22
No. 29	Algebra	Knowing	59	66	71
No. 30	Geometry	Using Routine Procedures	25	23	37
No. 31	Algebra	Knowing	33	43	57
No. 32	Fractions and Number Sense	Investigating and Solving Problems	34	39	45
No. 33	Data Representation, Analysis and Probability	Using Complex Procedures	31	31	79
No. 34	Algebra	Knowing	17	29	37
No. 35	Geometry	Using Complex Procedures	25	28	46
No. 36	Algebra	Knowing	32	40	47
Average			**34**	**37**	**52**

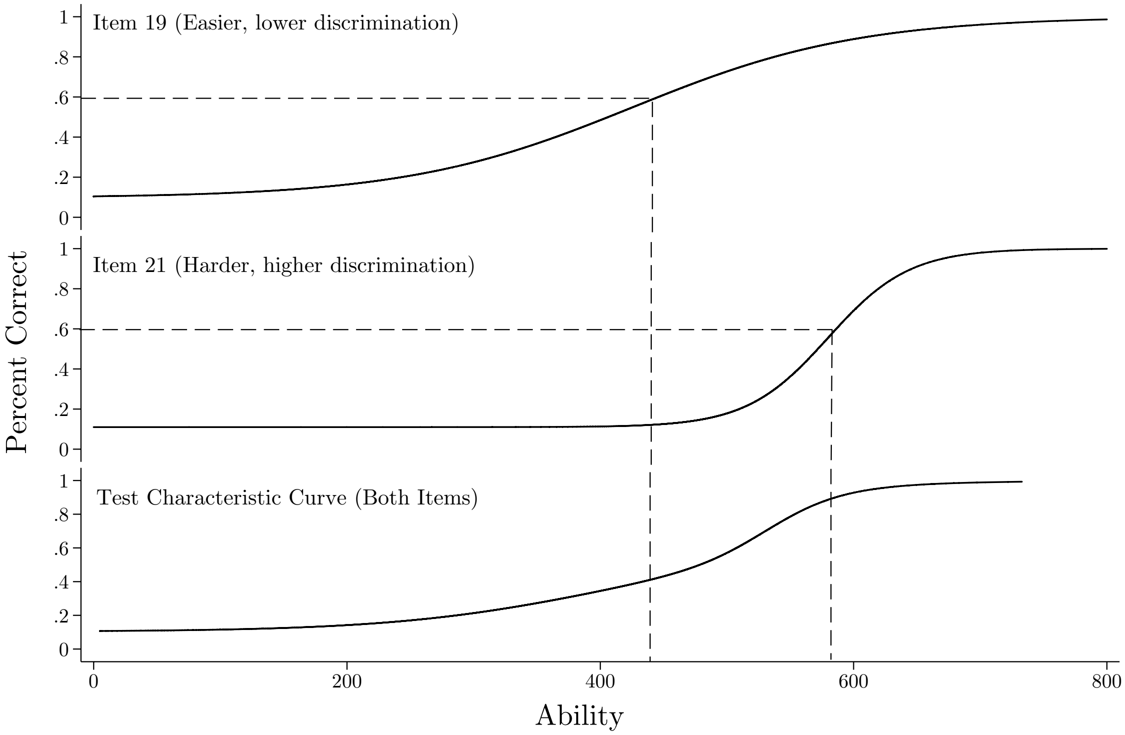
Source: This table from Wu et al (2006) summarizes the test results from the Rajasthan and Orissa Secondary School Survey, 2005 and TIMSS 1999.

Figure 1. Three-Parameter Logistic (3PL) Item Response Function



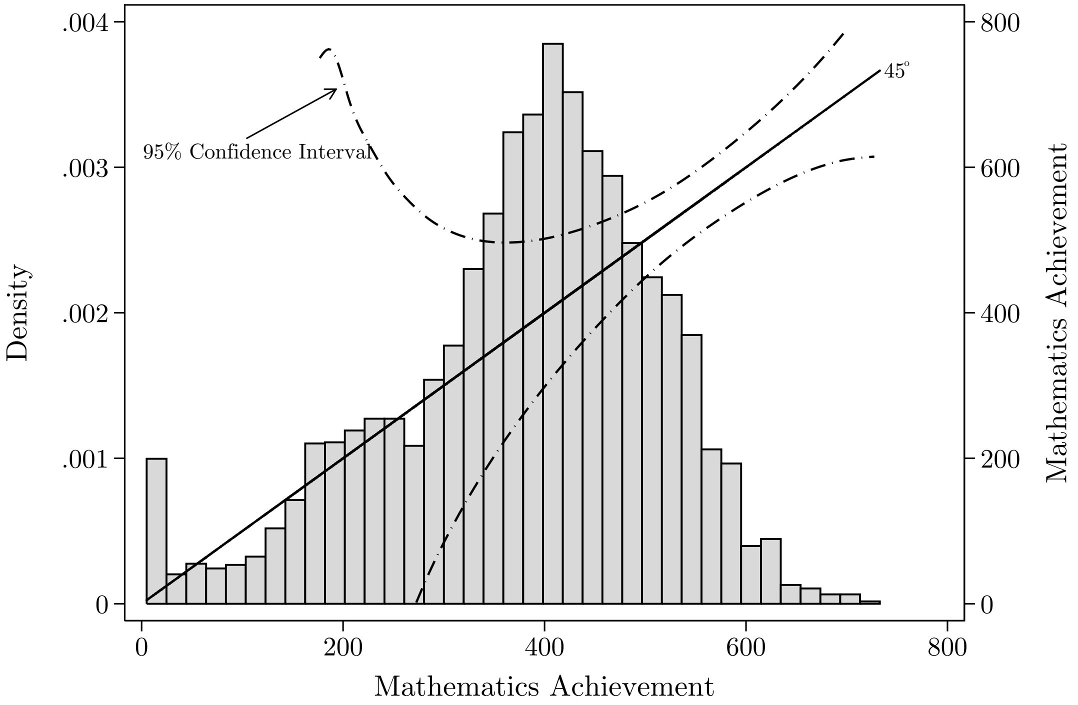
Notes: The parameters a, b and c represent the item discrimination, difficulty and pseudo guessing parameters, respectively.

FIGURE 2. RELATIONSHIP BETWEEN PERCENT CORRECT, ABILITY, AND THE TEST CHARACTERISTIC CURVE



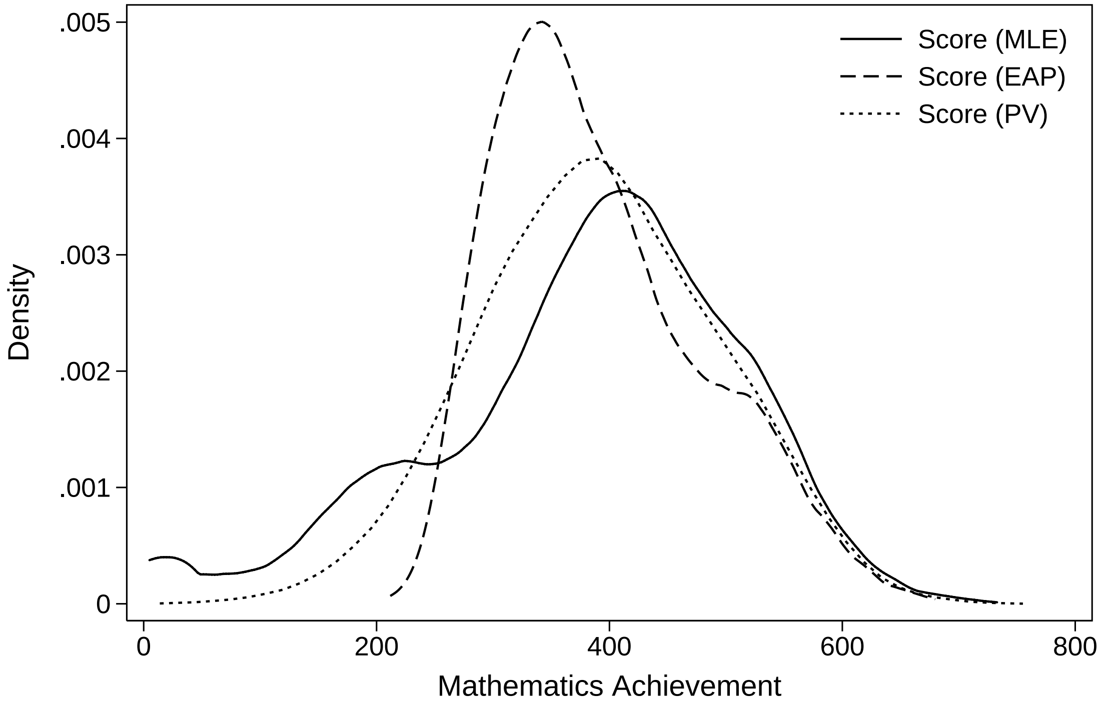
Notes: The top two panels give the item characteristic curves for items 19 and 21. The bottom panel shows the test characteristic curve of an exam which only presents these two items. One can read the link between the percent correct and latent ability using the x- and y-axes (dashed lines).

FIGURE 3. PRECISION OF MLE SCORES



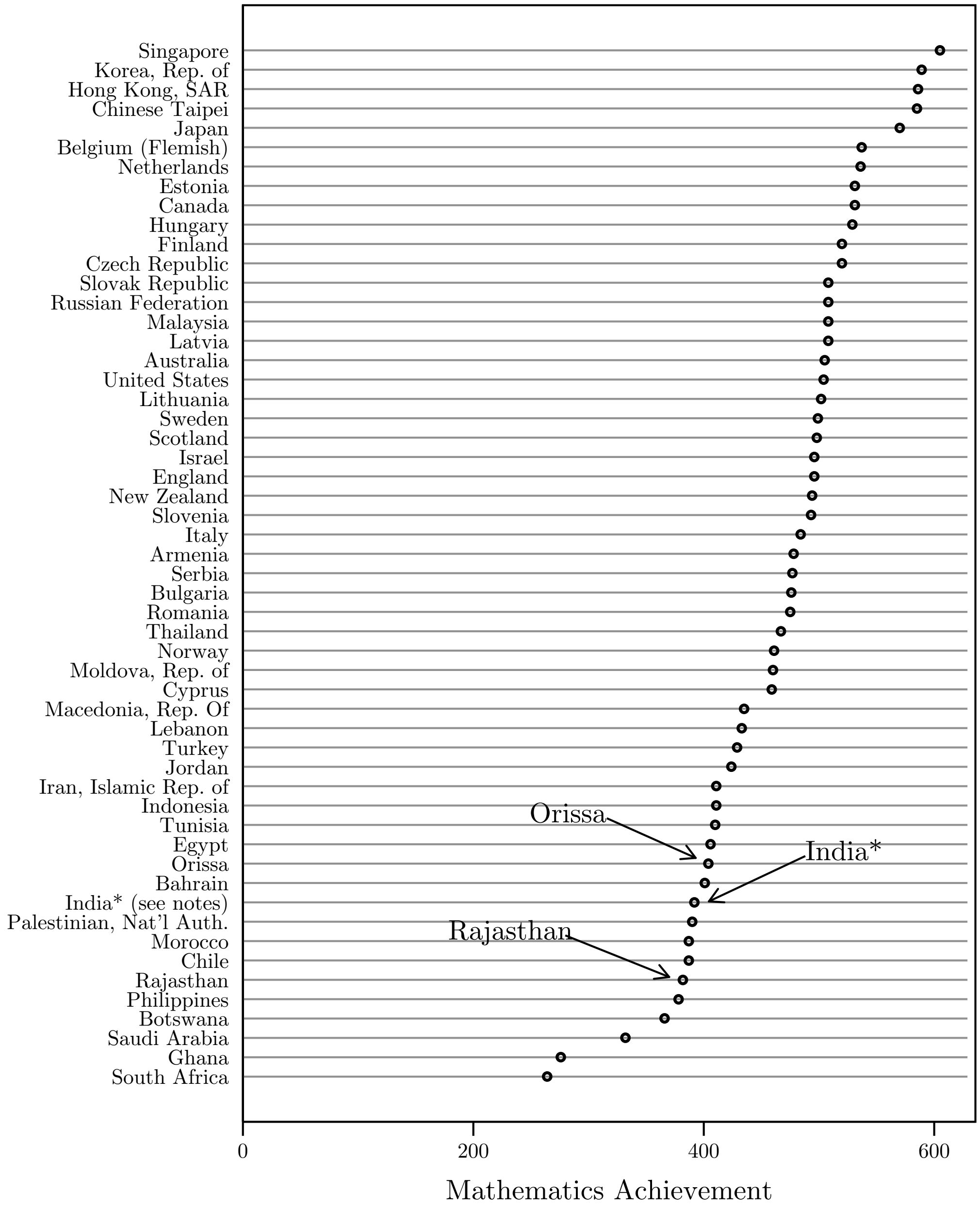
Notes: Plot of MLE achievement on MLE achievement (slope=1) with upper and lower 95% confidence intervals for individual scores plotted as dotted lines. The MLE score histogram (an overestimate of the true population distribution) is plotted in gray. MLE standard errors are computed from the inverse Fisher information matrix and thus graphed as symmetrical +/- 1.96*se. In reality, the precision is not symmetrical; large standard errors arise because it is difficult to discriminate between low and very low achievers, and between high and very high achievers.

Figure 4. Distribution of MLE, EAP and Plausible Value Scores



Notes: The MLE, EAP, and PV score distributions are represented by a kernel density. As discussed in the text, the true population distribution is bounded by the MLE and EAP estimates and given by the PV estimates. The PV kernel density was averaged over five plausible values per student. For reference, the average international score is 487 in 2001 and 467 in 2003.

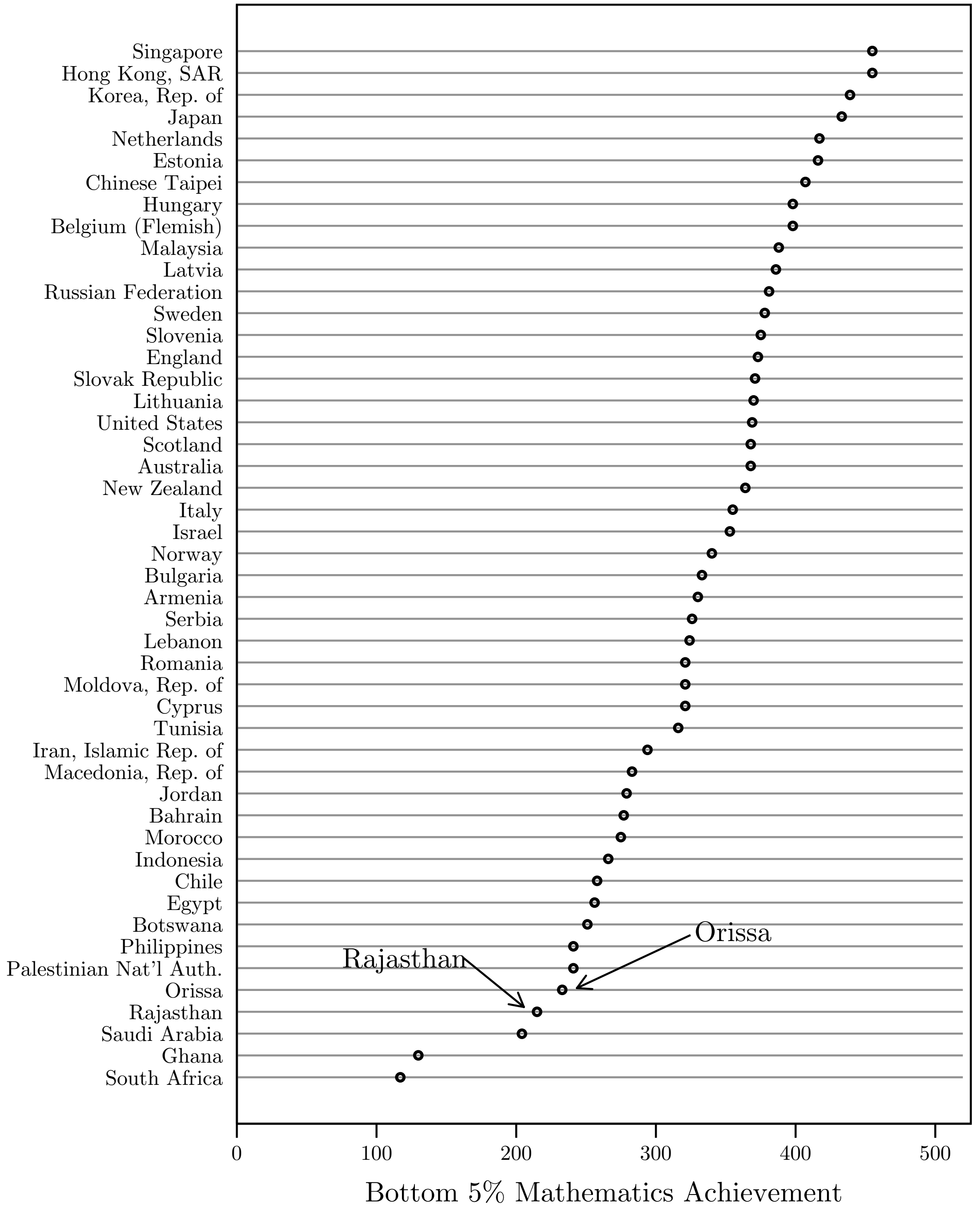
Figure 5. Average International Mathematics Achievement, Including Orissa and Rajasthan



Source: TIMSS Mathematics 2001, Grade 8, Exhibit 1.1; TIMSS Mathematics 2003, Grade 8, Exhibit D.1; authors' calculations from Rajasthan and Orissa Secondary School Survey, Grade 9, 2005. Most recent year is used when both are available.

Notes: *India is given purely as a best guess and assumes the tested sample within Orissa and Rajasthan represents India as a whole. We only report EAP estimates of the mean since these are most closely analogous to the TIMSS methodology.

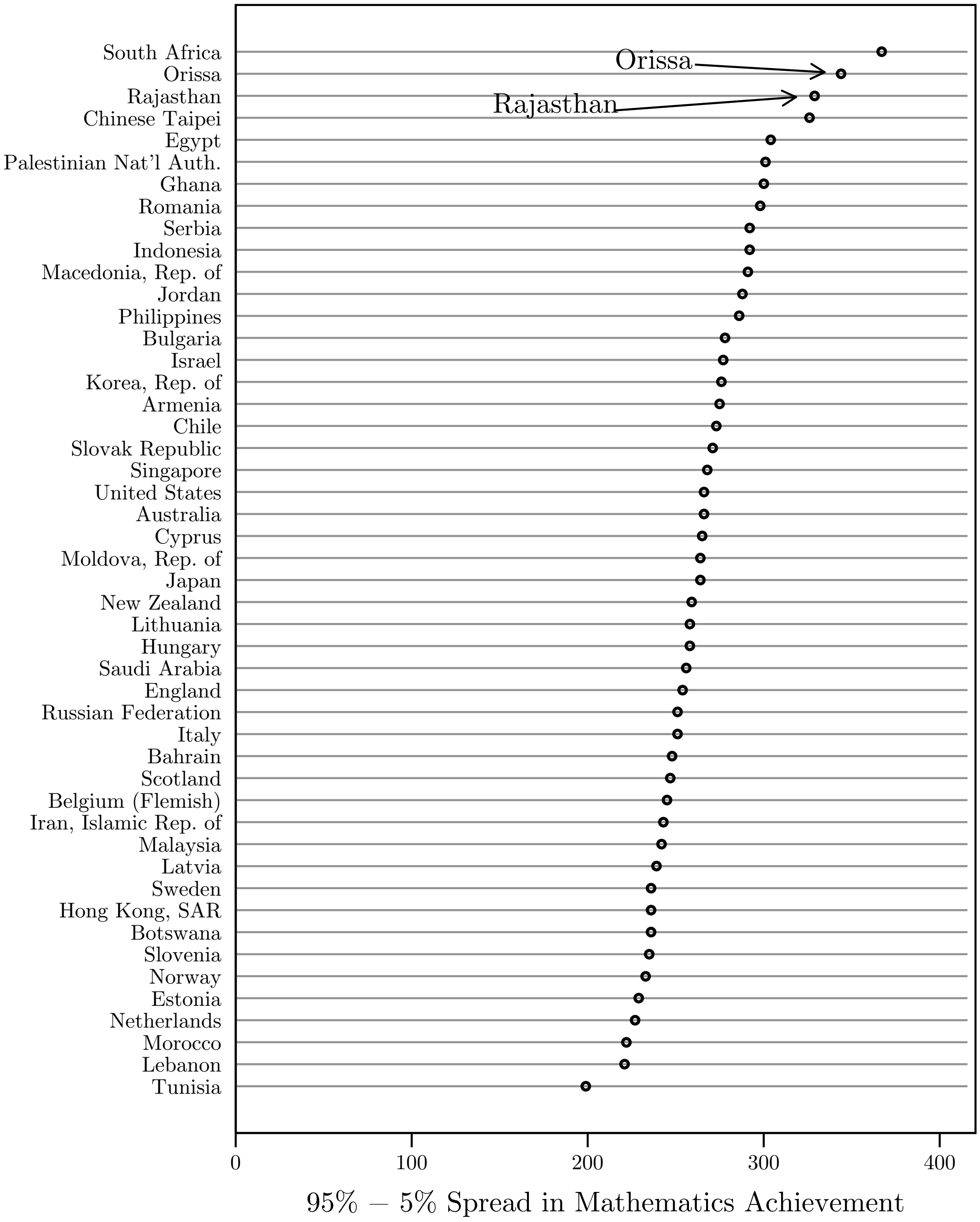
Figure 6. Bottom 5th Percentile of Mathematics Achievement, Including Orissa and Rajasthan



Source: TIMSS Mathematics 2001, Grade 8, Exhibit 1.1; TIMSS Mathematics 2003, Grade 8, Exhibit D.1; authors' calculations from Rajasthan and Orissa Secondary School Survey, Grade 9, 2005. Most recent year is used when both are available.

Notes: The Indian percentiles are calculated using the plausible values methodology described in the text.

Figure 7. 95th - 5th Percentile Spread of Mathematics Achievement, Including Orissa and Rajasthan



Source: TIMSS Mathematics 2001, Grade 8, Exhibit 1.1; TIMSS Mathematics 2003, Grade 8, Exhibit D.1; authors' calculations from Rajasthan and Orissa Secondary School Survey, Grade 9, 2005. Most recent year is used when both are available.

Notes: The Indian 95%-5% spread is calculated using the plausible values methodology described in the text.

Table 2. TIMSS 2003 International Benchmarks of Mathematics Achievement

Advanced International Benchmark – 625

Students can organize information, make generalizations, solve non-routine problems, and draw and justify conclusions from data. They can compute percent change and apply their knowledge of numeric and algebraic concepts and relationships to solve problems. Students can solve simultaneous linear equations and model simple situations algebraically. They can apply their knowledge of measurement and geometry in complex problem situations. They can interpret data from a variety of tables and graphs, including interpolation and extrapolation.

High International Benchmark – 550

Students can apply their understanding and knowledge in a wide variety of relatively complex situations. They can order, relate, and compute with fractions and decimals to solve word problems, operate with negative integers, and solve multi-step word problems involving proportions with whole numbers. Students can solve simple algebraic problems including evaluating expressions, solving simultaneous linear equations, and using a formula to determine the value of a variable. Students can find areas and volumes of simple geometric shapes and use knowledge of geometric properties to solve problems. They can solve probability problems and interpret data in a variety of graphs and tables.

Intermediate International Benchmark – 475

Students can apply basic mathematical knowledge in straightforward situations. They can add, subtract, or multiply to solve one-step word problems involving whole numbers and decimals. They can identify representations of common fractions and relative sizes of fractions. They understand simple algebraic relationships and solve linear equations with one variable. They demonstrate understanding of properties of triangles and basic geometric concepts including symmetry and rotation. They recognize basic notions of probability. They can read and interpret graphs, tables, maps, and scales.

Low International Benchmark – 400

Students have some basic mathematical knowledge.

Source: Exhibit 2.1, TIMSS 2003.

Table 3. Percent of Enrolled Children Passing International Mathematics Benchmarks, Including Orissa and Rajasthan

Country	***Low International Benchmark (>400)***	**Intermediate International Benchmark (>475)**	**High International Benchmark (>550)**	**Advanced International Benchmark (>625)**
Singapore	99	93	77	44
Korea, Rep. of	98	90	70	35
Hong Kong, SAR	98	93	73	31
Japan	98	88	62	24
Netherlands	97	80	44	10
Estonia	97	79	39	9
Chinese Taipei	96	85	66	38
Hungary	95	75	41	11
Belgium (Flemish)	95	82	47	9
Malaysia	93	66	30	6
Latvia	93	68	29	5
Russian Federation	92	66	30	6
Sweden	91	64	24	3
Slovak Republic	90	66	31	8
Australia	90	65	29	7
United States	90	64	29	7
Lithuania	90	63	28	5
Scotland	90	63	25	4
Slovenia	90	60	21	3
New Zealand	88	59	24	5
Israel	86	60	27	6
Italy	86	56	19	3
Bulgaria	82	51	19	3
Armenia	82	54	21	2
Norway	81	44	10	0
Serbia	80	52	21	4
Romania	79	52	21	4
Cyprus	77	45	13	1
Moldova, Rep. of	77	45	13	1
Lebanon	68	27	4	0
Macedonia, Rep. Of	66	34	9	1
Jordan	60	30	8	1
Indonesia	55	24	6	1
Iran, Islamic Rep. of	55	20	3	0

Tunisia	55	15	1	0
Egypt	52	24	6	1
Bahrain	51	17	2	0
Orissa	**50**	**27**	**9**	**1**
Palestinian, Nat'l Auth.	46	19	4	0
Rajasthan	**42**	**17**	**4**	**1**
Morocco	42	10	1	0
Chile	41	15	3	0
Philippines	39	14	3	0
Botswana	32	7	1	0
Saudi Arabia	19	3	0	0
South Africa	10	6	2	0
Ghana	9	2	0	0

Source: TIMSS Mathematics 2003, Grade 8, Exhibit 2.2 and authors' calculations from Rajasthan and Orissa Secondary School Survey, Grade 9, 2005.

Notes: Countries ranked by percent passing low benchmark. Estimates based on plausible values. All percentiles are for enrolled and tested children only.

Table 4. Estimated Number of 14-year Olds Passing International Mathematics Benchmarks, in Thousands

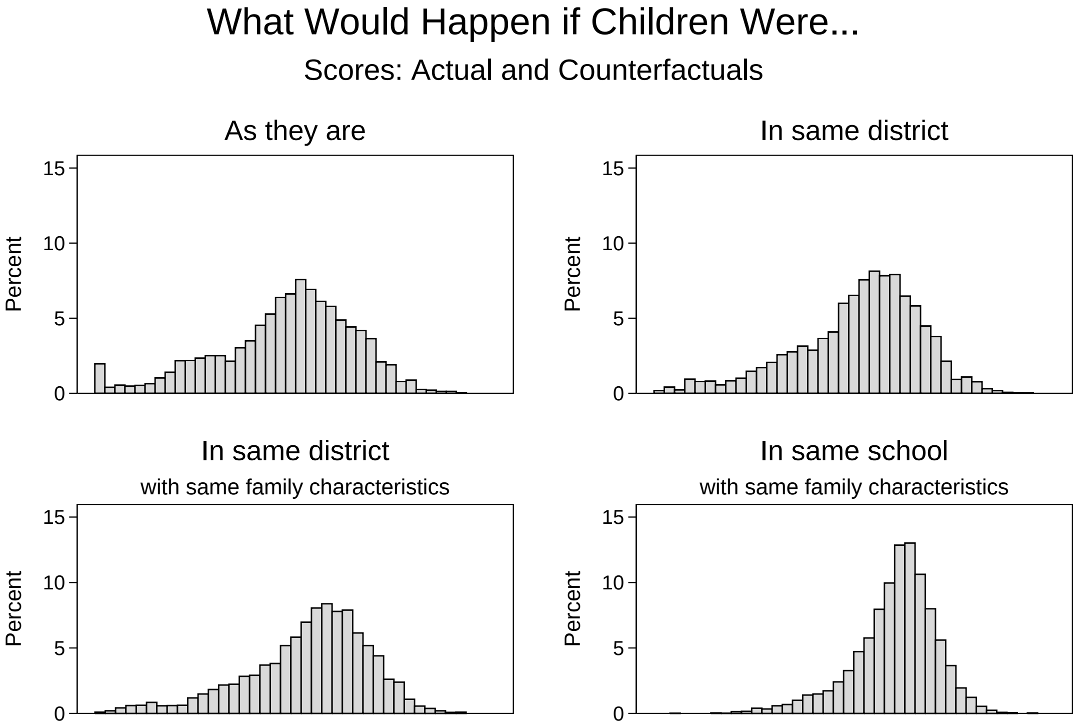
Country	**Not Enrolled or Below Low International Benchmark (<400)**	**Low International Benchmark (>400)**	**Intermediate International Benchmark (>475)**	**High International Benchmark (>550)**	***Advanced International Benchmark (>625)***
Japan	26	1189	1068	753	291
United States	792	3316	2358	1069	258
Korea, Rep. of	81	633	581	452	226
Chinese Taipei	28	291	257	200	115
India*	**17589**	**4634**	**2216**	**705**	**101**
Russian Federation	389	1080	775	352	70
Indonesia	3128	1424	622	155	26
Malaysia	155	363	258	117	23
Singapore	0	52	48	40	23
Hong Kong, SAR	19	59	56	44	19
Netherlands	27	174	144	79	18
Australia	64	215	155	69	17
Italy	112	436	284	96	15
Egypt	967	675	312	78	13
Hungary	16	100	79	43	12
Belgium (Flemish)	10	111	96	55	11
Romania	87	153	101	41	8
Israel	26	83	58	26	6
Serbia	52	82	53	21	4
Slovak Republic	18	50	37	17	4
Sweden	12	104	73	27	3
New Zealand	10	49	33	13	3
Bulgaria	20	52	33	12	2
Lithuania	7	38	26	12	2
Armenia	14	38	25	10	1
Jordan	70	66	33	9	1
Latvia	6	19	14	6	1
Estonia	2	13	10	5	1
Slovenia	3	17	11	4	1
Philippines	1528	478	172	37	0

Iran, Islamic Rep. of	787	592	215	32	0
South Africa	900	59	35	12	0
Chile	194	89	33	7	0
Norway	13	49	26	6	0
Moldova, Rep. of	25	38	22	6	0
Morocco	597	103	25	2	0
Lebanon	31	34	13	2	0
Macedonia, Rep. Of	14	16	8	2	0
Tunisia	120	70	19	1	0
Cyprus	3	9	5	1	0
Saudi Arabia	535	59	9	0	0
Ghana	521	18	4	0	0
Botswana	36	9	2	0	0
Bahrain	7	6	2	0	0

* We assume that the average passing rates for Rajasthan and Orissa approximates the percentage of children passing each benchmark for India as a whole.

Notes: Cells contain the estimated number of 14-year olds passing each benchmark, in thousands, based on Table 6, the net enrollment rate (WDI, 2005), and population data (U.S. Census Bureau, Population Division, International Programs Center, 2005). We assume children who are not enrolled do not pass the low-benchmark. Scotland and the Palestinian National Authority Territories were dropped for lack of population data. Enrollment rate is for the most recent reported years or imputed if only the gross rate is available.

Figure 8. Collapsing the Score Distribution



Notes: Panel 1 shows the distribution of MLE math scores for all children. We use MLE scores since fixed effects were not included in the construction of plausible values (see Mislevy et al, 1992). Panel 2 shows the residual distribution controlling for a district fixed effect. Panel 3 shows the residual distribution controlling for a district fixed effect and child age, age squared, gender, caste, mother literacy, father literacy, and household wealth. Panel 4 shows the residual distribution including an additional school fixed effect. A considerable portion of the distribution is Panel 4 is due to measurement error.

Table 5. Variance Decomposition: Districts, Households, Schools, and Children

Percent of variance explained by:	**Orissa**	**Rajasthan**
Same district	9	4
And household characteristics	6	4
And same school	32	41
Remaining (idiosyncratic)	53	52

Notes: Cells contain the percentage of variance (partial R-squared) explained by (a) a district fixed effect; (b) a district fixed effect and child age, age-squared, gender, caste, mother literacy, father literacy, and household wealth; and (c) a school fixed effect and child age, age squared, gender, caste, mother literacy, father literacy, and household wealth. The idiosyncratic variation includes any remaining variation, a substantial portion of which is measurement error.

Table A1. Average International Mathematics Achievement in 1999 and 2003, Including Orissa and Rajasthan

Country	Average Score 2003	Average Score 1999
Singapore	605 (3.6)	604 (6.3)
Korea, Rep. of	589 (2.2)	587 (2.0)
Hong Kong, SAR	586 (3.3)	582 (4.3)
Chinese Taipei	585 (4.6)	585 (4.0)
Japan	570 (2.1)	579 (1.7)
Belgium (Flemish)	537 (2.8)	558 (3.3)
Netherlands	536 (3.8)	540 (7.1)
Canada		531 (2.5)
Estonia	531 (3.0)	
Hungary	529 (3.2)	532 (3.7)
Finland		520 (2.7)
Czech Republic		520 (4.2)
Malaysia	508 (4.1)	519 (4.4)
Latvia	508 (3.2)	505 (3.4)
Russian Federation	508 (3.7)	526 (5.9)
Slovak Republic	508 (3.3)	534 (4.0)
Australia	505 (4.6)	525 (4.8)
United States	504 (3.3)	502 (4.0)
Lithuania	502 (2.5)	482 (4.3)
Sweden	499 (2.6)	
England		496 (4.1)
Scotland	498 (3.7)	
Israel	496 (3.4)	466 (3.9)
New Zealand	494 (5.3)	491 (5.2)
Slovenia	493 (2.2)	530 (2.8)
Italy	484 (3.2)	479 (3.8)
Armenia	478 (3.0)	
Serbia	477 (2.6)	
Bulgaria	476 (4.3)	511 (5.8)
Romania	475 (4.8)	472 (5.8)
Thailand		467 (5.1)
Norway	461 (2.5)	
Moldova, Rep. of	460 (4.0)	469 (3.9)
Cyprus	459 (1.7)	476 (1.8)
Macedonia, Rep. Of	435 (3.5)	447 (4.2)
Lebanon	433 (3.1)	
Turkey		429 (4.3)
Jordan	424 (4.1)	428 (3.6)

Iran, Islamic Rep. of	411 (2.4)	422 (3.4)
Indonesia	411 (4.8)	403 (4.9)
Tunisia	410 (2.2)	448 (2.4)
Egypt	406 (3.5)	
Orissa	**404 (1.7)**	
Bahrain	401 (1.7)	
India* (see notes)	**392 (1.1)**	
Palestinian, Nat'l Auth.	390 (3.1)	
Chile	387 (3.3)	392 (4.4)
Morocco	387 (2.5)	337 (2.6)
Rajasthan	**382 (1.4)**	
Philippines	378 (5.2)	345 (6.0)
Botswana	366 (2.6)	
Saudi Arabia	332 (4.6)	
Ghana	276 (4.7)	
South Africa	264 (5.5)	275 (6.8)

Source: TIMSS Mathematics 2001, Grade 8, Exhibit 1.1; TIMSS Mathematics 2003, Grade 8, Exhibit D.1; authors' calculations from Rajasthan and Orissa Secondary School Survey, Grade 9, 2005.

Notes: *India is given purely as a best guess and assumes the tested sample within Orissa and Rajasthan represents India as a whole. We only report EAP estimates of the mean since these are most closely analogous to the TIMSS methodology. Parentheses contain standard errors of the mean.

Table A2. Distribution of International Mathematics Achievement in 2003, Including Orissa and Rajasthan

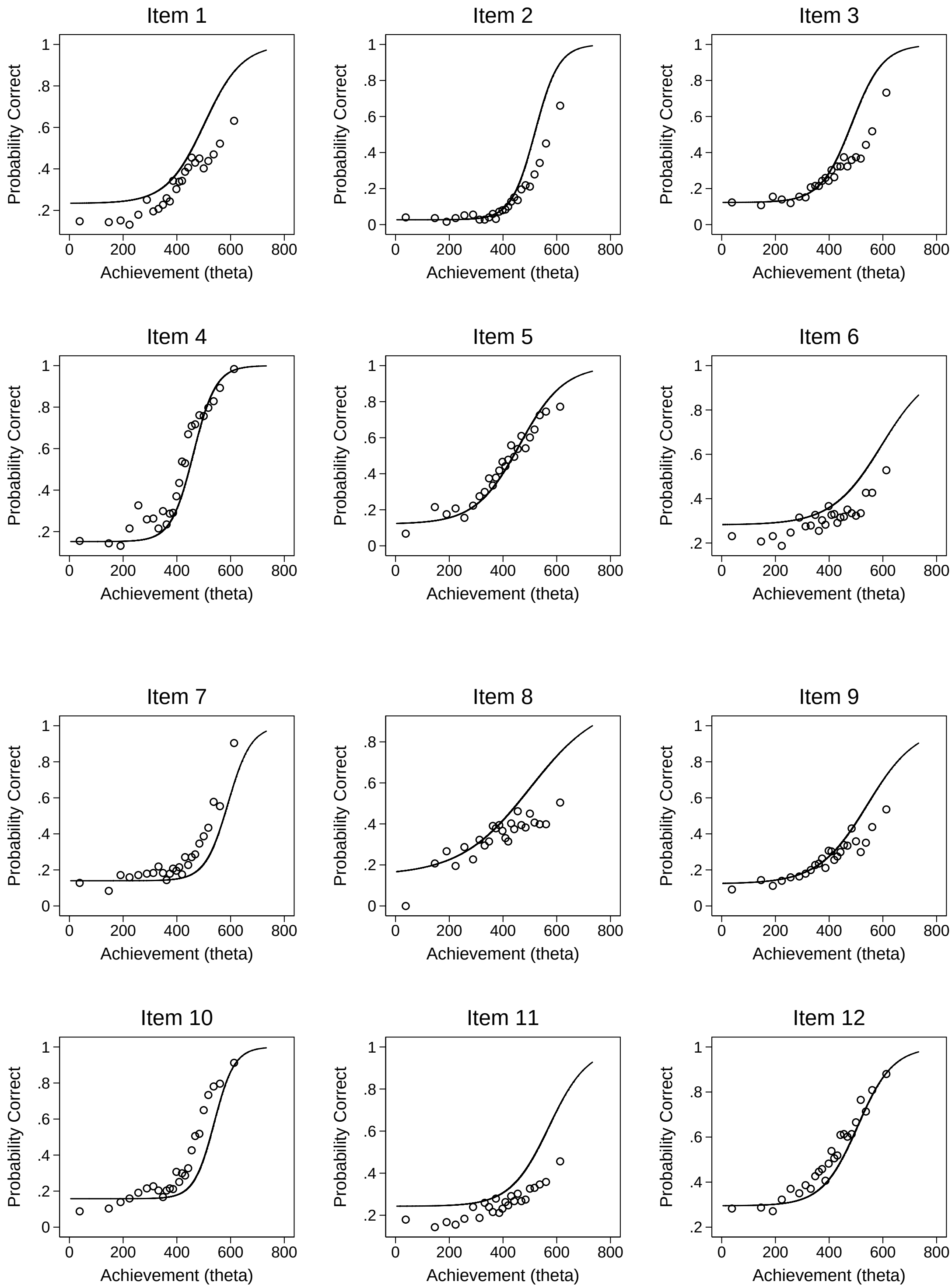
Country	**5th percentile**	**25th percentile**	***50th percentile***	**75th percentile**	**95th percentile**
Singapore	455(6.6)	556(6.7)	614(4.0)	662(3.5)	723(2.8)
Chinese Taipei	407(6.0)	518(7.0)	596(4.6)	657(5.0)	733(6.0)
Korea, Rep. of	439(3.1)	537(3.2)	596(2.5)	647(2.5)	715(3.0)
Hong Kong, SAR	455(11.9)	546(4.0)	593(3.3)	635(3.0)	691(4.6)
Japan	433(4.4)	519(2.0)	572(2.6)	623(2.2)	697(5.1)
Belgium (Flemish)	398(8.9)	495(3.7)	545(3.1)	588(2.8)	643(3.3)
Netherlands	417(8.4)	488(4.5)	540(5.8)	587(4.8)	644(6.8)
Estonia	416(4.8)	484(3.6)	531(4.0)	577(2.7)	645(4.0)
Hungary	398(8.1)	476(2.9)	531(3.5)	584(4.1)	656(4.2)
Latvia	386(5.2)	458(5.2)	510(2.9)	559(3.5)	625(5.4)
Russian Federation	381(5.5)	456(4.2)	509(4.5)	561(4.0)	632(7.5)
Slovak Republic	371(6.5)	453(4.7)	509(3.9)	564(4.3)	642(4.2)
Malaysia	388(3.7)	455(3.9)	507(5.5)	562(6.1)	630(5.3)
Australia	368(10.4)	450(3.9)	506(3.7)	561(5.8)	634(6.6)
United States	369(4.7)	450(2.9)	505(3.0)	560(3.5)	635(3.8)
Lithuania	370(4.5)	448(2.9)	503(2.4)	557(4.0)	628(2.5)
Scotland	368(8.5)	449(5.0)	501(4.3)	550(3.9)	615(6.0)
Sweden	378(4.0)	452(4.3)	501(2.6)	548(2.9)	614(6.3)
Israel	353(5.9)	438(4.8)	498(5.3)	555(3.5)	630(5.3)
England	373(5.3)	445(5.9)	497(5.9)	552(9.2)	627(5.6)
New Zealand	364(9.9)	441(5.2)	495(5.3)	548(7.1)	623(12.5)
Slovenia	375(9.3)	445(2.4)	492(2.0)	542(1.6)	610(3.7)
Italy	355(6.0)	432(4.0)	486(2.9)	537(3.2)	606(5.0)
Armenia	330(7.5)	423(5.1)	483(3.3)	539(3.2)	605(3.5)
Romania	321(7.8)	413(4.6)	479(4.9)	540(4.9)	619(9.0)
Serbia	326(6.2)	417(4.8)	479(4.0)	540(3.1)	618(4.8)
Bulgaria	333(7.5)	421(5.5)	478(4.6)	535(4.6)	611(6.6)
Norway	340(5.2)	414(2.2)	465(3.3)	511(1.7)	573(2.4)
Moldova, Rep. of	321(5.8)	405(7.3)	464(4.9)	518(4.4)	585(5.1)
Cyprus	321(3.8)	405(3.4)	463(1.8)	518(1.5)	586(1.6)
Macedonia, Rep. of	283(4.8)	376(5.1)	439(2.9)	497(3.4)	574(4.7)
Lebanon	324(3.4)	387(3.9)	432(3.7)	479(4.0)	545(5.8)
Jordan	279(5.3)	362(4.1)	427(4.9)	488(5.0)	567(5.2)
Indonesia	266(11.6)	350(7.9)	411(6.0)	472(4.0)	558(3.6)
Iran, Islamic Rep. of	294(4.8)	360(3.5)	408(3.0)	461(2.4)	537(6.2)
Tunisia	316(2.2)	368(2.4)	407(2.4)	450(2.6)	515(6.2)
Egypt	256(3.0)	341(6.0)	405(4.1)	471(3.7)	560(3.2)
Bahrain	277(3.2)	347(1.5)	402(1.8)	455(2.2)	525(1.4)

Orissa	**233**	**329**	**401**	**482**	**577**
Palestinian Nat'l Auth.	241(5.2)	326(3.2)	389(4.1)	455(4.2)	542(5.4)
Morocco	275(4.8)	340(3.0)	387(3.0)	434(3.0)	497(2.8)
Chile	258(4.5)	328(4.8)	382(3.4)	441(4.5)	531(4.9)
Rajasthan	**215**	**312**	**381**	**449**	**544**
Philippines	241(3.6)	316(5.6)	373(6.4)	437(6.5)	527(8.0)
Botswana	251(5.1)	316(3.0)	365(2.5)	415(2.7)	487(5.0)
Saudi Arabia	204(10.0)	279(6.6)	331(5.1)	385(4.5)	460(5.4)
Ghana	130(5.8)	213(4.3)	274(5.3)	337(7.3)	430(9.1)
South Africa	117(5.2)	191(3.5)	248(4.0)	316(7.5)	484(20.1)

Source: TIMSS Mathematics 2003, Grade 8, Exhibit D.1 and authors' calculations from Rajasthan and Orissa Secondary School Survey, Grade 9, 2005.

Notes: Countries ranked by median score. Estimates of population percentiles computed using plausible values. Parentheses contain standard errors.

Figure A1 – TIMSS Item Response Functions and Observed Responses



Item 1

Item 2

Item 3

Item 4

Item 5

Item 6

Item 7

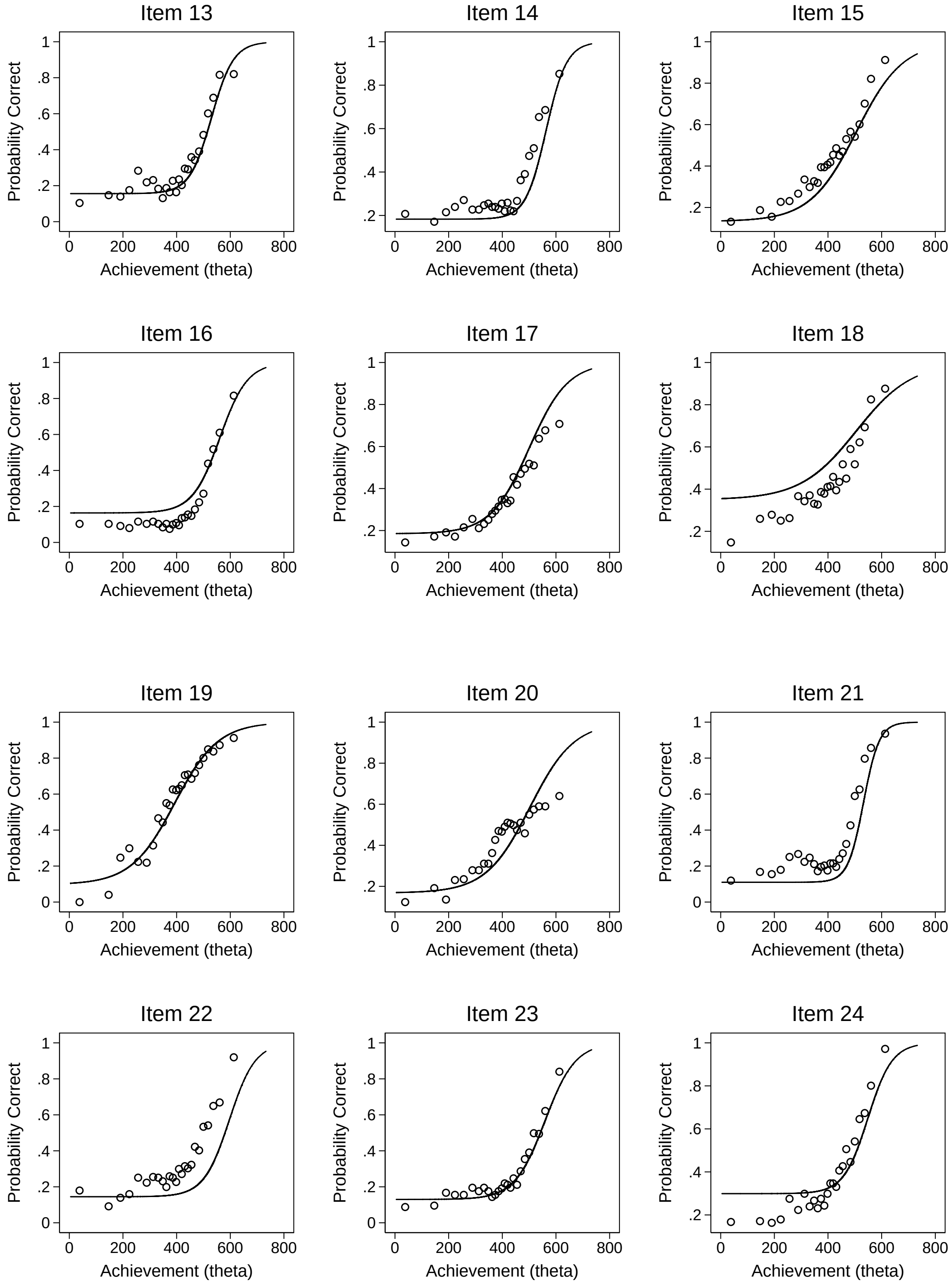
Item 8

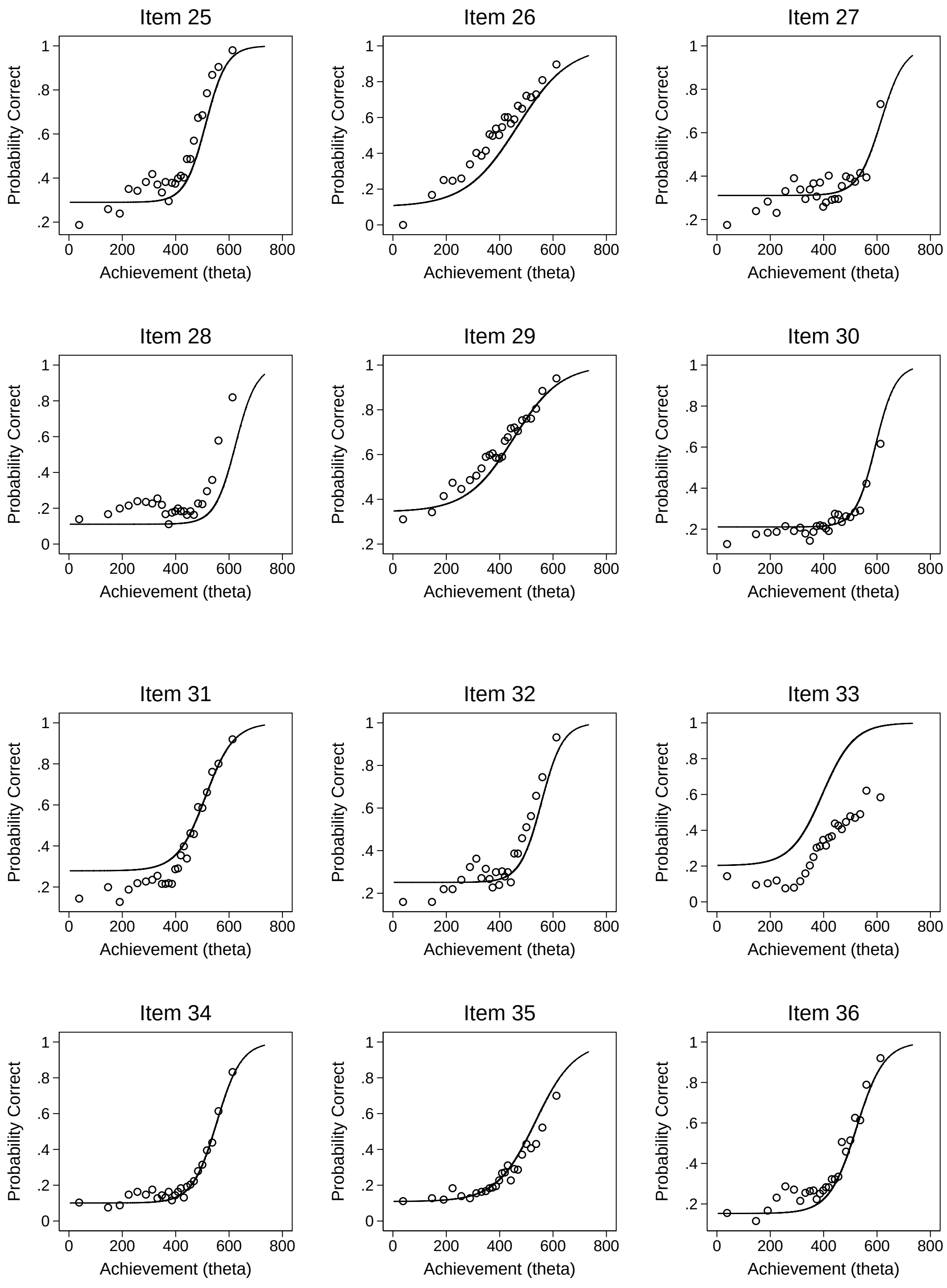
Item 9

Item 10

Item 11

Item 12


Item 13
Item 14
Item 15
Item 16
Item 17
Item 18
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Probability Correct
Achievement (theta)
0
200
400
600
800
0
.2
.4
.6
.8
1



Notes: Observed responses (dots) are means of 25 achievement bins. Expected responses (lines) use fixed TIMSS item parameters.