Has India's learning crisis really worsened?

Sarthak Joshi*

October 2, 2024

Abstract

Despite near-universal enrolment and continuing progress across multiple input-based measures of learning, standardized test scores amongst Indian children show a large and permanent decline after 2010. I argue that this puzzling decline is partly an artefact of changing measurement error in the main source of learning outcome data in the country. Using an external benchmark, I show that pre-2011 estimates are systematically biased upwards. Bias disappeared and data quality improved after robust survey procedures were introduced in 2011. Even if the real decline was smaller than previously thought, concerns around declining productivity of the Indian education system remain.

JEL Classification: I20, I21

Keywords: Education in India, learning metrics, survey measurement error

*Email: sarthak.joshi@warwick.ac.uk. I am especially grateful to Abhijeet Singh for his feedback and support at multiple stages of this project. I thank Andreas Stegmann and Sonia Bhalotra for valuable comments.

1 Introduction

Despite significant expansions in enrollment and grade progression in recent decades, learning levels amongst children living in many developing countries remain persistently low (World Bank, 2018). India exemplifies this. In 2018, half of grade five students in rural India couldn't read a text at the grade two level and two-thirds could not solve a three-by-one division problem (Pratham, 2019).

Learning levels are not just low, but also appear to have declined in the last decade, possibly quite dramatically after 2010 (Shah and Steinberg, 2019). While the evidence for persistently low learning levels is strong,[1] a large and sudden decline would pose a substantial puzzle - and a challenge - for policymakers and researchers on education in India. This is because indicators of household wealth, levels of school attendance, and most input-based measures of school productivity like per-pupil expenditure, pupil-teacher ratios, and physical infrastructure have improved during this period (Muralidharan, 2019). Low-cost private schooling has also seen impressive expansion in the last two decades, significantly expanding parents' choice of schooling in rural areas.[2]

Any explanation for the decline in aggregate learning levels must contend with three facts observed in the main source of data on learning trends in the country - the Annual State of Education Report (ASER).[3] First, the decline is entirely concentrated between 2010 and 2012. Learning levels were relatively stable both before and after this two-year period. Second, the magnitude of the decline is large (0.3 sd for math scores and 0.2 sd for reading scores), and equally apparent at all points of the learning distribution. Third, the drop in learning levels observed between 2010 and 2012 has been remarkably persistent. Despite some improvement between 2016 and 2018, learning levels in 2018 were lower than 2008 levels in most states of India. Any valid explanation for the learning decline must therefore be powerful enough to have plausibly caused a large and permanent drop in learning levels over just 2 years.

The most common such explanation has to do with the national roll out of the Right to Education (RTE)[4] Act starting from the 2010 academic session (Shah and Steinberg, 2019; Muralidharan,

[1] Nationally representative surveys, as well as smaller, more detailed studies (Muralidharan et al., 2019; Rossiter et al., 2018) have consistently found that learning outcomes amongst Indian children are much lower than grade-appropriate levels.

[2] The proportion of children ages 6-14 attending a private school in rural India increased from 22.5% in 2008 to 30.8% in 2018 (Pratham, 2019). Compared to government schools, private schools in rural India typically deliver similar or slightly better learning outcomes at a substantially lower cost per student (Muralidharan and Sundararaman, 2015; Kingdon, 2020).

[3] Launched in 2005 by the NGO Pratham, the ASER survey was amongst the first nationally representative of learning outcomes in the country, and has since played a central role in bringing attention to the low levels of learning amongst Indian children (Bridgespan, 2018). It remains the only such survey that has been conducted regularly using a consistent set of learning assessments.

[4] The RTE was a wide-ranging educational reform which mandated (i) free and compulsory education for all children

2019; Pratham, 2019). Taking the ASER data series as given, these explanations typically point to the role played by the RTE in (i) relaxing classroom teaching due to the introduction of automatic academic promotion until grade 8, (ii) increasing load on teachers due to the introduction of a system of continuous evaluations, and (iii) worsening peer effects due to an influx of previously out-of-school students into the schooling system to explain why learning levels may have declined. While plausible, empirical evidence in support for the RTE-based explanation is mixed. For instance, exploiting geographical variation induced by the introduction of the 'No Detention Policy' as part of the RTE, Ahsan et al. (2018) find that the policy led to a significant *improvement* in reading and math scores. RTE-based explanations also fail to account for the fact that states that experienced large declines in the last decade (like Madhya Pradesh, Chhattisgarh, and Maharashtra) begin to show a declining trend before the RTE came into effect.

In this paper, I offer a new explanation for the learning decline by taking a closer look at the data itself. ASER introduced robust data quality assurance procedures during the 2011 survey round,[5] the same year in which aggregate learning levels appear to fall for the first time. I argue that these improvement in survey procedures led to a reduction in systematically positive non-sampling survey measurement error over time, contributing to the appearance of a learning decline in the data. I present two pieces of evidence that support this interpretation.

First, I conduct an external benchmarking exercise using data from the India Human Development Survey (IHDS) at two points in time - once before ASER introduced robust survey procedures (in 2006/7), and once after (in 2012). The IHDS is a high-quality nationally representative survey which employs a similar sampling strategy and uses exactly the same assessment tool as ASER. As such, it serves as a valid external benchmark. For most states, I find that estimates of average test scores for the same underlying population of children (rural; ages 8-11) in ASER's 2007 survey round are systematically higher than estimates in the 2006 round of the IHDS. However, estimates across the two surveys agree closely in 2012.[6] Since learning levels in ASER data were very stable in the years between 2007-2010 and 2012-2016, I take this as evidence that ASER's estimates are biased upwards throughout the pre-2011 era, but not in surveys conducted after 2011.

ages 6 to 14, (ii) minimum norms around school infrastructure, (iii) a 25 percent reservation for disadvantaged students in private schools, and (iv) automatic academic promotions and scrapping of high-stakes year-end exams for students until class 8.

[5] While survey procedures have continuously improved since ASER launched its first survey in 2005, changes introduced in 2011 were significantly wider in scope and incorporated real-time monitoring of data collection for the first time. See section 5 for details.

[6] The correlation of state-level learning estimates across IHDS and ASER data increased from only 0.45 for math scores and 0.58 for reading scores in 2006/7 to 0.92 in math scores and 0.79 in reading scores in 2012.

Second, I document the fact that ASER introduced significant improvements in data quality assurance procedures during the 2011 survey round and provide evidence to show that these improvements in survey procedures translated into better data quality in post-2011 ASER survey data. I exploit the district-level panel structure of ASER data to decompose the variance of aggregate learning estimates into persistent ('signal') and transitory ('noise') components using variance decomposition techniques outlined in Kane and Staiger (2002a,b). Defining 'noise' as the share of variance that is due to transitory factors (arising from non-sampling survey measurement error), I find that estimates of aggregate test scores in ASER survey data are significantly less noisy in the period after the introduction of robust data quality assurance procedures (2012-2018; roughly 5% of variance is noise) than before (2007-2011; roughly 20% of variance is noise).

Taken together, my findings indicate that the real decline in learning outcomes amongst Indian children between 2007 and 2012 was likely of a smaller magnitude than previously thought. Perversely, this is only because ASER surveys conducted before 2011 were overestimating learning outcomes - real learning levels have been low all along. My findings thus provide further support to the stylized fact that learning levels amongst Indian children are persistently low (Muralidharan, 2019). By reducing the magnitude of decline that requires explanation, my findings deliver a partial solution to the puzzle of declining learning outcomes after 2010 despite consistent improvements in input-based measures of learning (Shah and Steinberg, 2019).

Finally, my results contribute to the growing evidence base on the lack of agreement across different sources of learning outcome data in India (Bansal and Roy, 2019). Using direct audit evidence, Singh (2020) found that official learning metrics in the state of Madhya Pradesh are severely inflated due to data manipulation. Johnson and Parrado (2021) find that state rankings based on the government-run National Achievement Survey (NAS) display almost no correlation with state rankings based on ASER, IHDS, or net state domestic product per capita. In general, this body of evidence underlines the need for investing in reliable and high-frequency learning metrics to correctly diagnose and solve the learning crisis unfolding in developing countries (Pritchett, 2015; Figlio et al., 2016).[7] My results indicate that estimates of learning levels constructed using ASER survey data collected before 2012 are positively biased and contain relatively high levels of non-sampling survey measurement error. Previous work that has used information on learning levels in

[7] This is more important than ever given that the learning crisis has likely worsened since 2020 due to long school closures during the COVID-19 pandemic. Early evidence points to large learning losses and increasing inequalities due to school closures, even in some of the richest parts of the world (Engzell et al., 2021). Schools in low- and middle-income countries have remained closed up to twice as long as schools in high-income countries on average (UIS, 2022).

ASER data between 2007-2011, or on changes that include the period between 2010 and 2012 may need to be re-examined in light of my findings.[8]

In the absence of identifying variation in survey data quality over time, I am unable to provide a correction for actual trends in the ASER data series before 2012. Nonetheless, it is clear that rapid improvements in input-based measures of learning did not translate to improved learning outcomes during this time.[9] Understanding the determinants of this decline in productivity of the Indian education system and designing cost-effective policies to improve quality at scale remains the central challenge facing Indian policymakers.

The rest of the paper is organized as follows. Section 2 describes data sources. Section 3 describes trends in learning outcomes in ASER survey data between 2007-2018. Section 4 presents results from an external benchmark exercise using IHDS data to show that ASER's estimates are biased upwards in 2006/7, but not in 2012. Section 5 presents a variance decomposition exercise to show that data quality appears to have improved after robust new procedures were introduced during the 2011 survey round. Section 6 provides a discussion of the results. Section 7 concludes.

2 Data

The main source of data used in this paper are the ASER surveys. Conducted by Pratham, one of India's largest NGOs, ASER is a home-based survey that collects information on basic literacy and numeracy for a nationally representative sample of 5-16-year-olds across rural India. Data are available at an annual frequency between 2007 and 2014 and for every other year between 2014 and 2018. Nearly all rural districts are covered in each survey year. The survey employs a two-stage sampling design, with villages being sampled from each district in the first stage (using probability proportional to size sampling) and households in the second stage. All children aged 5-16 in selected households are tested. Estimates are representative at the district level.[10] In 2018, ASER was conducted in 596 out of 640 districts of India and tested 390,830 children across 354,944 households.

The second source of learning outcome data used in this paper are the India Human Development

[8] A search for "ASER data"+"India" on Google Scholar returned 441 results in August 2022.

[9] The government-run National Achievement Survey (NAS) also finds a small decline in test scores amongst students in grade 5 between cycle 3 (2010) and cycle 4 (2014). Note that the magnitude of change in the NAS is not directly comparable to either the IHDS or ASER since the NAS employs a different sampling strategy and assessment tool than ASER/IHDS. See table A1 in the online appendix for a comparative description of ASER, IHDS, and NAS.

[10] In 2016, ASER's sampling frame was updated from the list of 593 districts as per the 2001 census to the list of 640 districts as per the 2011 census. I construct a panel of districts with unchanged boundaries between 2007 and 2018 using the 2001 sampling frame.

Surveys (IHDS). Run jointly by the University of Maryland and the National Council of Applied Economic Research (NCAER), the IHDS is a nationally representative multi-topic panel survey of households with data available at two points in time - 2005/6 and 2011/12. The IHDS tests all children aged 8-11 in sampled households using ASER's assessment tool and employs a two-stage sampling strategy that is very similar to ASER.[11] As such, it is possible to construct directly comparable aggregate measures of learning levels across the IHDS and ASER surveys (at the state level) after applying appropriate restrictions.

ASER's learning assessment tool is designed to measure foundational reading and arithmetic skills. Every child, regardless of age or enrollment status, is asked four questions on basic maths and reading, respectively. The four maths questions ask whether the child can recognize numbers 1–9, recognize numbers 10–99, subtract two two-digit numbers with borrowing, and divide a three-digit number with a one-digit number. The four questions on reading ability are administered in the child's native language, and check whether the child can recognize letters, recognize words, read a short paragraph and read a short story in the local language. The score for each question is coded as 1 if the child correctly answers the question and 0 otherwise. A sample of ASER's math and reading assessment tools is shown in figure A1. Unless otherwise specified, I make use of a cumulative measure of learning proficiency for each child constructed in the following manner. A "math score" corresponds to the simple sum of the scores across the four numeracy questions asked in the ASER survey. For example, if a child correctly recognizes numbers between 1–9 and 10–99, and correctly answers the subtraction question but cannot correctly answer the division question, then that child's math score is coded as 3. A "reading score" is constructed in a similar manner. Every child aged between 5-16 years in the data is thus assigned a math score and a reading score, each ranging from 0 to 4.

3 Learning outcome trends in ASER data

3.1 Trends in average learning levels

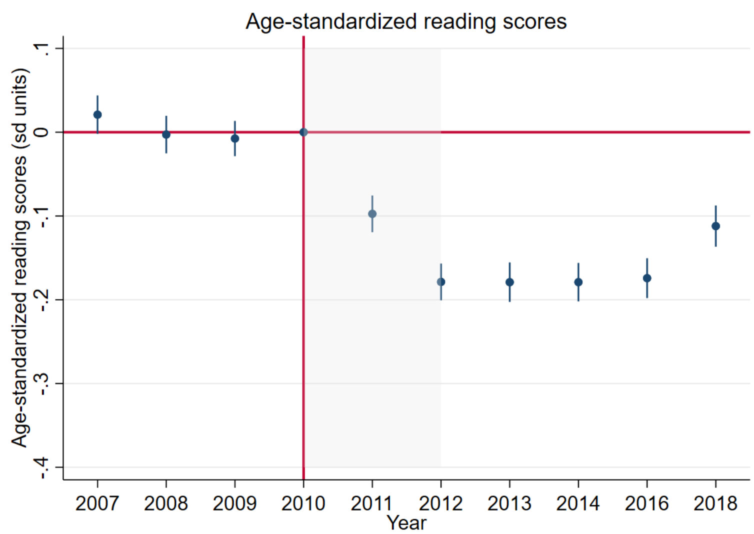
To investigate how average learning levels change over time in ASER data, I follow the event study-style approach outlined in Shah and Steinberg (2019) to estimate the following model using child-level data collected between 2007 and 2018 for all children ages 5-16:

[11] The IHDS uses the 2001 census sampling frame and covers 384 out of 593 districts across both rural and urban areas of India. During the 2011/12 round, the rural sample of the IHDS consisted of 27,579 households across 1,420 villages. Attrition in rural areas is very low - 90% of all households were re-interviewed in 2011/12.

$$S_{ijt} = \alpha + \beta_t + \gamma \cdot \mathbf{X}_i + \delta_j + \epsilon_{ijt}, \quad (1)$$

where S_{ijt} is the math or reading score attained by child i in district j in year t. Scores have been normalized with respect to age-specific means and standard deviations of national-level scores in 2018. This scaling implies that coefficients can be interpreted in terms of age-independent standard deviation units with respect to 2018 data. β_t is a vector of the coefficients for each survey year, $\mathbf{X}_i$ includes controls for child age and sex, and δ_j is a vector of district fixed effects. Figure 1 graphs β_t from regression 1 for math and reading scores for each survey year. Regressions are clustered at the district level, and 95 per cent confidence intervals are shown as bars in the figure. The omitted year is 2010.

Figure 1: Math and Reading scores in ASER data



Note: This figure shows estimates of β_t from equation 1 along with 95 per cent confidence intervals. The base year is 2010.

Figure 1 shows that the observed decline in learning outcomes in ASER data between 2007 and 2018 is entirely accounted for by steep year-on-year declines between 2010 and 2012. Scores were remarkably stable outside of this two-year period,[12] although there were some signs of improvement between 2016 and 2018. The magnitude of the decline was large - average scores decreased by 0.3 sd units for math and 0.2 sd units for reading during this time. Figure A2 shows trends in aggregate test scores by state. Dropping states that experienced the largest declines between 2010 and 2012

[12] Shah and Steinberg (2019) find similar patterns using raw test scores.

leaves the general pattern of learning decline unchanged, which suggests that it is not driven by outliers.

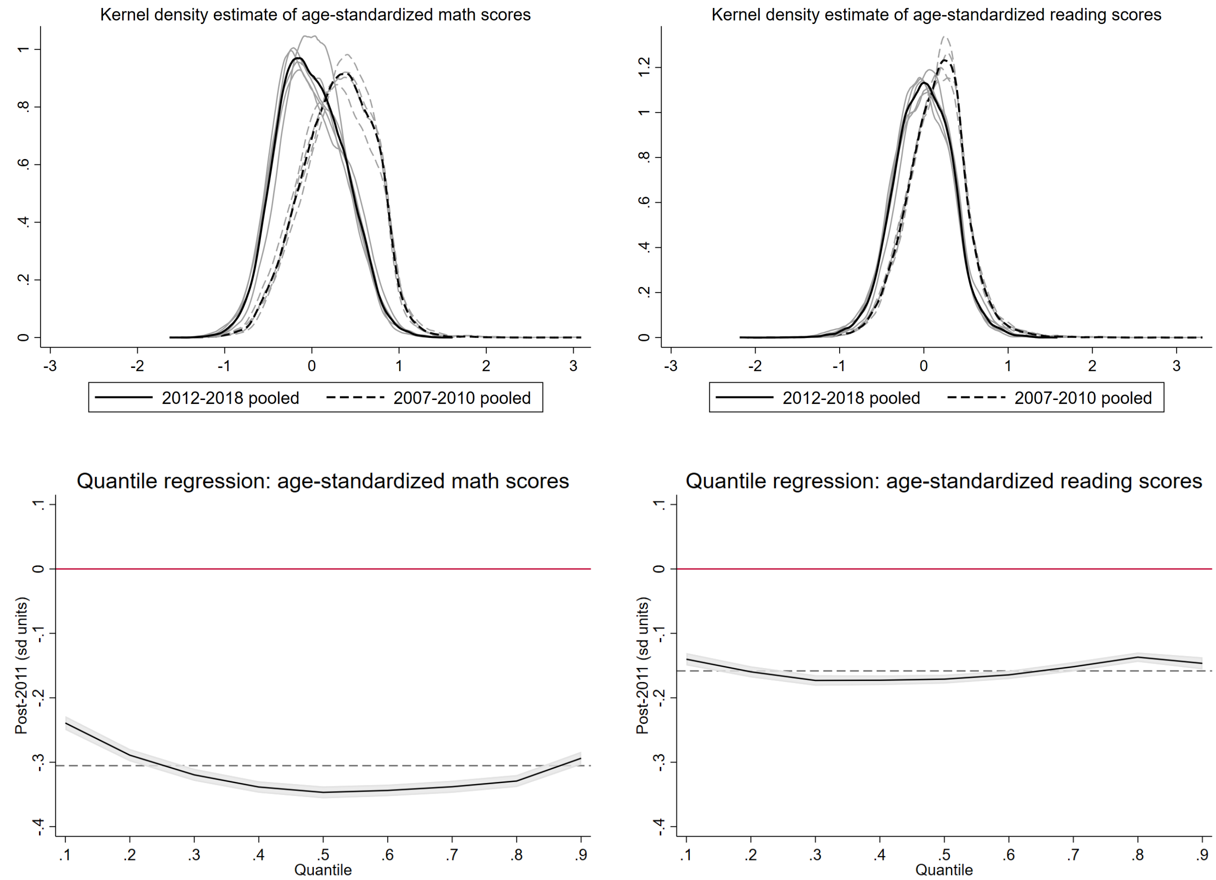
3.2 Looking across the learning distribution

The top panel of figure 2 shows kernel density estimate plots of the distribution of test scores aggregated at the district level across all survey years. Distributions constructed using data from surveys conducted between 2007-2011 and 2012-2018 appear to bunch up closely together, with the pooled distribution (shown in thick black lines) for the latter lying almost entirely to the right of the former. This reaffirms the finding from the previous section that the decline was entirely concentrated between 2010 and 2012, with learning levels being remarkably stable outside of this two-year period.

The bottom panel of figure 2 presents estimates from a quantile regression in which age-standardized scores were regressed on a dummy variable that takes value one for data collected between 2012-2018 and zero for data collected before 2011. This analysis shows that the decline took place at every point of the learning distribution by roughly the same (large) magnitude.

A decline of this magnitude in the national average over just two years, despite near-universal rates of school attendance and consistent improvements in relevant characteristics like household wealth and mother's education (see figure A3) represents a substantial negative shock to basic math and reading ability. For comparison, consider that Das et al. (2020) find that a large earthquake in Northern Pakistan that left more than 80,000 people dead and virtually all physical infrastructure destroyed in 2005, led to a relative decline in learning outcomes of *only* 0.24 SD amongst children living close to affected regions, four years after the incident.

Figure 2: Shift in the distribution of age-standardized math and reading scores



Note: The top panel shows kernel density estimate plots of district-level aggregates of test scores constructed using survey data from 2007-2010 (dashed lines) and 2012-2018 (solid lines). Distributions of pooled data are in black. Individual survey years are in light grey. Data from 2011 has been excluded to aid interpretation. The bottom panel shows results from a quantile regression where scores were regressed on a dummy variable that takes value one for data collected between 2012-2018, and zero otherwise. The dashed line shows the corresponding OLS estimate.

The analysis presented in this section has established three facts about the observed decline in learning outcomes in ASER data between 2007-2018. First, the decline was entirely concentrated between 2010 and 2012. Second, the magnitude of the decline was large (0.3 sd for math scores and 0.2 sd for reading scores), and equally apparent at all points of the learning distribution. Third, the decline has been remarkably persistent. Despite some improvement between 2016 and 2018, learning levels in 2018 remained lower than 2008 levels in most states of India.

4 Are pre-2011 ASER estimates biased upwards?

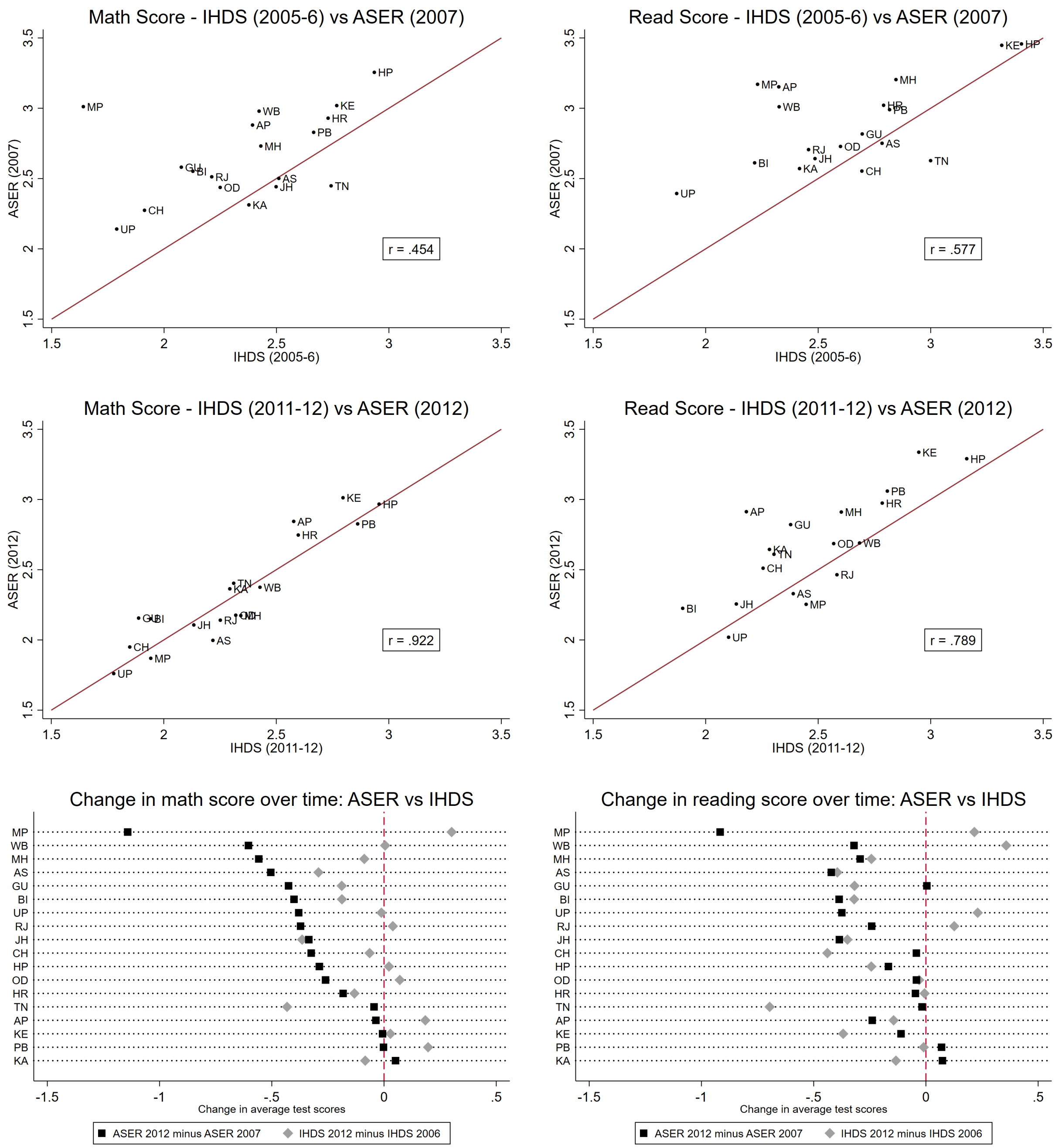
As mentioned before, the IHDS employs a similar sampling strategy and uses the same learning assessment tool as ASER. Besides being directly comparable to ASER surveys in these regards, there are at least two reasons to believe that the IHDS contains smaller non-sampling measurement error relative to ASER. First, ASER is implemented with the help of various partner organizations which in turn often use volunteer enumerators with relatively little experience. On the other hand, IHDS employs full-time survey teams that are put through a more extensive training period than is needed for single-topic surveys like ASER.[13] Second, while the IHDS randomly selects households in each village following a full household listing exercise, ASER employs a relatively less rigorous (although more cost-effective) right-hand rule for sampling households in sampled villages. Data from the IHDS can thus serve as a valid external benchmark for ASER data.

I make three adjustments to make the estimates across the IHDS and ASER directly comparable. First, I restrict the IHDS round I and round II samples to rural households only (ASER covers only rural areas). Second, I restrict the ASER samples to only include children aged between 8-11 years of age (the learning assessment tool in the IHDS is administered only to this age group). Finally, I apply appropriate sampling weights to ensure that estimates from each survey are representative at the state level.

The top two panels of figure 3 present a comparison of state-level estimates of raw math and reading scores across the two datasets. This comparison allows benchmarking of ASER estimates at two points in time - (i) once before learning levels appear to start falling in the ASER data (i.e. IHDS 2006 vs ASER 2007), and (ii) once after they appear to stabilize (i.e. IHDS 2012 vs ASER 2012). The bottom panel shows how test scores have changed over time in each data series.

[13] IHDS surveyors attend a two-week long training session followed by field practice sessions (see here), while ASER's surveyor training typically lasts for 3 days.

Figure 3: Comparing state-level math scores in ASER and IHDS across 2006/07 and 2012



Notes: The top two panels of this figure show a comparison of levels of raw math and reading scores at the state level across ASER and IHDS in 2006/7 and 2012. IHDS scores have been plotted on the x-axis and ASER scores on the y-axis. Points closer to the 45-degree reference line signal greater agreement between the two datasets. The bottom panel compares changes in raw test scores between 2006/7 and 2012 in ASER data (black squares) and IHDS data (grey diamonds).

The top panel of figure 3 shows that ASER's estimates of (raw) math and reading scores are systematically higher than IHDS's estimates for the same underlying population of children across most states in 2006/7 (r=0.45 for math and r=0.58 for reading). However, the middle panel shows that estimates agree very closely in 2012 (r=0.92 for math and r=0.79 for reading).[14] It follows that changes in test scores in ASER data (black squares) are systematically more negative than changes in IHDS data (grey diamonds) over the same time period, especially for maths (bottom panel of figure 3).

The two surveys thus paint a very different picture of how learning outcomes evolved between 2006/7 and 2012 amongst children aged 8-11 in rural India. While ASER data shows that age-standardized reading scores fell by 0.21 sd units (se=0.02) and math scores fell by 0.35 sd units (se=.02) during this time, IHDS data show a test scores remained largely stable, with math scores declining by only 0.01 sd units (se=0.03) and reading scores declining by only 0.05 sd units (se=0.02).[15]

To the extent that estimates of state-level math and reading scores constructed using IHDS data can serve as a valid external benchmark, this analysis suggests that learning levels in ASER data from 2007 are systematically biased upwards. Estimates constructed using 2012 ASER data, on the other hand, appear to be free from bias. Given that learning levels in ASER data were very stable in the years between 2007-2010 and 2012-2016, I take this as evidence for the fact that ASER's estimates are biased upwards throughout surveys conducted in the pre-2011 era, but not in surveys conducted after 2011.

Looking at how learning levels evolved in particular states can provide further clues. Consider the case of Madhya Pradesh (MP), the state that shows the largest disagreement across the two surveys in 2006/7. MP is India's fifth-largest state, with an estimated population of 85 million people in 2020. It is also amongst the poorest, with a nominal per-capita GDP of around \$1,400 in 2019, compared to the national average of \$2,100 in the same year. As per the 2011 national census, it had one of the lowest adult literacy rates in the country, at only 70%. Even so, MP was reported to have amongst the highest learning levels in the country in ASER's 2008 report, with 87% of grade 5 students being able to read a grade 2 level text that year (Pratham, 2009).[16] This was then reported to have dropped precipitously to 55% in 2010 and then to only 27% by 2012. Section B of the online appendix provides details on how learning levels evolved for stable cohorts of children in

[14] States that show the largest discrepancy in math scores in 2006/7 - Madhya Pradesh ("MP"), West Bengal ("WB"), and Andhra Pradesh ("AP") - also show large differences in reading scores. This may be evidence of generally high survey measurement error at the state level.

[15] In urban areas, IHDS data show that scores in both math and reading declined by only 0.06 sd units (se=0.03) between 2006/7 and 2012.

[16] In the same year, the corresponding share in Kerala and Himachal Pradesh - two states widely considered to have the highest quality of schooling provision in the country - was only 75.5% and 75.7%, respectively.

MP during this time. If ASER's data reflect reality, the implication from this cohort-level analysis is that children across *every age group* in MP forgot how to do basic arithmetic, enough to push them back to levels they had achieved 2-3 years earlier. While true learning loss is a possibility, I find the large magnitude and general nature of the decline (which began before the RTE was rolled out in 2010) implausible. On the other hand, my hypothesis that the decline in learning levels is an artefact of a reduction in positive bias is in line with the evidence.

5 What explains the reduction in bias?

In this section, I document the fact that ASER introduced robust survey protocols during the 2011 survey round, the same year in which aggregate learning levels appeared to decline for the first time. I then construct a measure of non-sampling measurement error and use it to show that surveys conducted after 2011 had significantly lower levels of noise compared to surveys conducted before 2011. I hypothesize that the same improvements can also plausibly account for why positive bias declined after 2011.

5.1 Changes in ASER's data quality assurance procedures

Over the years, ASER has instituted increasingly robust surveyor training and survey data quality assurance processes to ensure high-quality data collection. I document how these protocols have evolved over time using information available in publicly available annual reports and other survey documentation.[17]

Early ASER survey rounds employed only basic data quality assurance procedures. In 2008, the ASER centre was established as the dedicated research and assessment arm of Pratham and the ASER survey was brought within its ambit. Later in the same year, field re-checks were introduced for the first time. This entailed pre-selecting four out of the 20 villages selected in each district to be re-surveyed after the completion of the main survey. If survey quality was found to be lacking, the whole district was surveyed again. The next big set of improvements in the scope and depth of survey data quality assurance procedures came in 2011. Some key changes introduced during the 2011 survey round include:

- **Real-time monitoring:** Call centres were set up in each state to keep in touch with field supervisors to ensure compliance with processes of training, survey, and recheck. An SMS-

[17] ASER's data quality control framework is available here.

based recheck system was also introduced under which compiled district-level data was texted to a central server to enable immediate review. Field monitoring was also introduced for the first time in 2011.

- **Expanded re-checks:** Master Trainers and ASER state-level teams started conducting field rechecks. District-level desk and phone rechecks were introduced. Villages to be monitored and rechecked were selected based on predefined criteria starting in 2011. This ensured that poorly surveyed villages could be identified and resurveyed immediately. Between 2008 and 2011, villages to be rechecked were selected randomly.

- **More time spent in each district**: Before 2011, survey teams spent one weekend in each district before moving on to the next. Starting in 2011, however, the survey was conducted over two consecutive weekends in each district. While the total time spent in each village remained the same, spreading the survey over two weekends may have improved the quality of monitoring and re-checking by field supervisors.

- **Increase in training duration:** District-level surveyor training was extended from two to three days.

Many of the improvements in survey procedures introduced in 2011 were retained and strengthened further in future survey rounds.[18] Combined with the fact that data quality would have likely improved over time anyway as the ASER centre and its partner organizations gained more experience of survey implementation, it is likely that bias stemming from non-sampling measurement error in post-2011 ASER surveys would have been smaller than in earlier survey rounds.

5.2 Evidence for improved data quality after 2011

I adopt variance-decomposition methods outlined by Kane and Staiger (2002a)[19] to construct a measure of non-sampling survey measurement error in ASER survey data across two time periods - pre-2011 and post-2011. The basic idea is to construct a measure of the signal-to-noise ratio in

[18] SMS monitoring was expanded to all states by 2013. Starting in 2014, a village-wise database of monitored and re-checked villages was maintained to track data quality and focus on districts that required further attention. An external field re-check was also conducted for the first time in 9 states.

[19] I follow Johnson and Parrado (2021) in adopting this method. My approach differs from theirs on at least two counts. First, while the general focus of their analysis is to assess the internal reliability of ASER data collected between 2006-2014, my focus is to compare data quality before and after 2011. Second, while they use district-level estimates for children in grades 3, 4 and 5 (as reported in ASER reports), I make use of micro survey data collected between 2007-2018 for all children ages 5-16.

estimates of aggregate test scores in the ASER series while exploiting the panel structure of the data, but without explicitly modelling individual components that make up each component of variance in test scores.

Suppose that test scores are composed of two parts. One part represents 'persistent' factors influencing test scores, such as the quality of teachers, school infrastructure and the curriculum being used, which are highly correlated from year to year. The second part is made up of purely non-persistent factors that influence test scores, such as survey sampling error, measurement error, and any other transient factors that can influence district-level test scores in a given year like bad weather on the day of the test, short-term changes in district-level education funding etc. When comparing correlations between current scores and previous scores for increasing lags, one would expect correlation to fall by a large magnitude with the first lag (reflecting changes in both persistent and transitory components) and then exhibit relatively smaller declines for lags longer than one (reflecting changes largely in persistent components). Since persistent factors are slow-moving, one would also expect correlations with lags longer than one to decline at a steady rate.

Figure 4 shows that the data corresponds well with our theoretical expectations. For both reading and math scores, the initial decrease in correlation is larger than the subsequent decreases. Moreover, subsequent correlations decline at a relatively low and stable rate of 11 percent for math scores and 8 percent for reading scores for each additional lag.

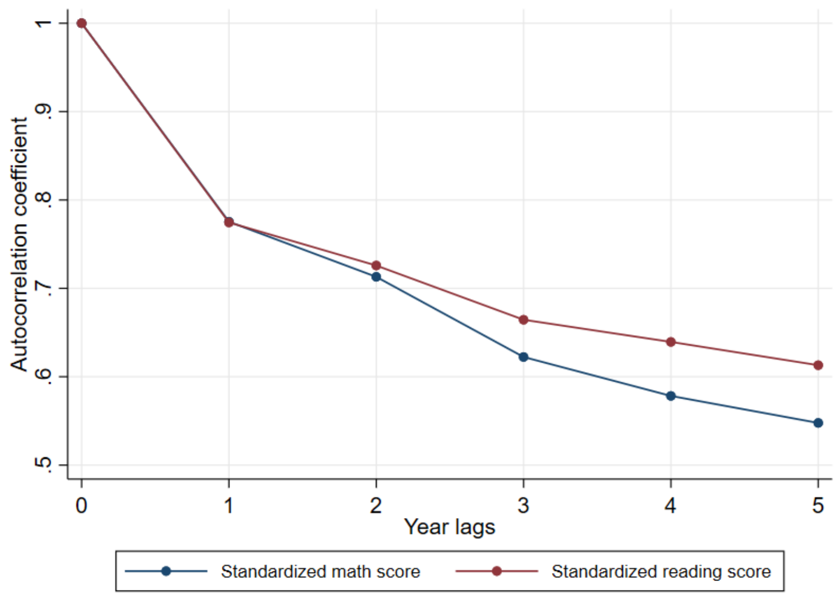
This way of interpreting the data suggests a method to decompose the first-order autocorrelation coefficient of test scores into persistent and non-persistent components. Begin by constructing an estimate of the purely persistent component of the first-order autocorrelation of test scores (ρ_{pers}) by taking an average of the proportional decline in correlations between lags 2 and 5:

$$\rho_{pers} = \frac{1}{4} \sum_{n=2}^{5} \frac{\rho_k}{\rho_{k-1}} \tag{2}$$

If only persistent factors were in play, one would expect test scores one year apart to also be correlated by ρ_{pers} on average. Any *additional* decline in the first-order autocorrelation coefficient can be interpreted as being caused by changes in non-persistent components. We can thus estimate the share of the variance due to persistent factors by taking the ratio of the actual correlation coefficient (ρ_1) and the expected correlation coefficient (ρ_{pers}). Multiplying this share with the total variance of test scores (σ_y^2) gives us the amount of persistent variance in test score data:

$$\sigma_{pers}^2 = \sigma_y^2 \cdot \frac{\rho_1}{\rho_{pers}} \tag{3}$$

Figure 4: Decay in autocorrelation by year lags in ASER data



Notes: This figure plots autocorrelation coefficients for math and reading test scores using district-level panel data from ASER surveys between 2012 and 2018.

Any residual variance in learning scores can then be interpreted as being driven by non-persistent or transient factors ($\sigma_{tran} = \sigma_y^2 - \sigma_{pers}^2$). Finally, subtracting away the component of variance arising due to survey sampling design[20] from this estimate of transient variance gives a measure of variance due to other (non-sampling) transient factors, which is my measure of 'noise':

$$\sigma_{tran,other}^2 = \sigma_y^2 - \sigma_{pers}^2 - \sigma_{tran,sampling}^2 \tag{4}$$

Using this method, I exploit the district-level panel structure of the data to decompose the variance of district-level estimates of math and reading scores into three sources - (i) persistent factors, (ii) sampling variance (transient), and (iii) other non-sampling transient factors. Results of this decomposition analysis conducted separately for three different time periods - 2007-2011, 2012-2018, and overall - have been presented in table 1.

[20] Sampling variance for each district was calculated by dividing the within-district variance by the district's sample size. The total variance due to sampling variance ($\sigma_{tran,sampling}^2$) is calculated as the average of sampling variances across all districts. Sampling variance is very small in ASER data.

Table 1: Share of persistent and non-persistent components in variance of learning levels

Years	Age-standardized math score				Age-standardized reading score			
	σ^2_y	σ^2_{pers}	$\sigma^2_{tran,other}$	noise share	σ^2_y	σ^2_{pers}	$\sigma^2_{tran,other}$	noise share
2007-2011	0.744	0.586	0.157	21.1%	0.721	0.573	0.147	20.4%
2012-2018	0.692	0.659	0.032	4.6%	0.762	0.707	0.054	7.1%
Overall	0.824	0.710	0.113	13.8%	0.771	0.646	0.124	16.1%

Notes: The value of $\sigma^2_{tran,other}$ is calculated using equation 4. The value of $\sigma^2_{tran,sampling}$ (not reported in table 1) is very small, varying between 0.0008 and 0.0012 across years and subjects.

Table 1 shows that estimates of learning levels constructed using survey data collected before the introduction of robust survey procedures in 2011 contain significantly higher non-sampling survey measurement error compared to estimates constructed using post-2011 survey data. This suggests that quality assurance procedures introduced during the 2011 survey round were successful in improving survey data quality. The analysis presented in section C in the online appendix shows that a similar conclusion holds when looking at changes (rather than levels).

6 Discussion

I provided evidence to show that pre-2011 estimates of learning levels in ASER data are systematically biased upward, but post-2011 estimates are free from bias. I argued that the introduction of new survey procedures in 2011, which seem to have reduced noise, can also plausibly explain why bias declined after 2011.

This still leaves the question of why pre-2011 estimates in ASER data are systematically biased upwards. While it is not possible to provide a definitive answer to this question in the absence of identifying variation in survey data quality over time, one hypothesis relates to how surveyors were administering the ASER test in the early years of the survey. For instance, it is possible that enumerator training workshops held in certain states (like Madhya Pradesh) failed to effectively train enumerators to not prompt or help children while solving learning tasks. Surveyors may have also marked students at higher levels due to fear of scrutiny. Knowledge about the extent of

India's learning crisis was still in its infancy during the early ASER rounds, and surveyors may have avoided marking children at levels which were too far below age-appropriate learning levels to avoid being pulled up by supervisors. Finally, artificially higher scores in pre-2011 surveys may be related to how long it takes to administer the ASER tool to students at different levels of learning. In particular, the ASER learning tool is laid out so that students who end up with a score between 2-4 answer exactly two questions, but those who end up with a score of 1 answer exactly three questions (see figure A1). As such, it would take somewhat longer to complete the survey in villages where many children are at the lowest level of learning. In the absence of real-time monitoring of data quality in the pre-2011 era, it is possible that enumerators were neglecting to mark children at lower levels of learning to ensure that survey deadlines were being met.

7 Conclusion

Despite near-universal rates of primary school enrollment and continuous improvement in input-based measures of schooling, the main source of data on learning trends in India - the ASER surveys - show that basic arithmetic and reading ability amongst children aged 5-16 declined significantly between 2007 and 2016. The observed decline in learning levels was large (0.3 sd for math and 0.2 sd for reading), rapid (entirely concentrated between 2010 and 2012), highly persistent (levels in 2018 had still not recovered to pre-2010 levels), and visible at all points of the learning distribution.

Previous work has typically pointed to changes in policy like the national rollout of the RTE in 2010 to explain the learning decline. I offer a new explanation by taking a closer look at the data itself. ASER introduced robust data quality assurance procedures in 2011, the same year in which aggregate learning levels appeared to decline for the first time. I show that estimates of learning levels in pre-2011 ASER surveys are systematically biased upwards and contain a relatively high share of non-sampling measurement error. On the other hand, estimates constructed using post-2011 survey data are free from this positive bias, and significantly less noisy. Based on this evidence, I argue that the observed decline in learning levels between 2007-2012 is plausibly an artefact of a reduction in positive bias due to improved data quality over time. Even if the real decline was of a smaller magnitude than previously thought, concerns around the declining productivity of the Indian education system remain.

In general, my findings serve as a reminder for researchers to pay careful attention to measurement issues in survey data, while underlining the importance of investing in reliable and high-frequency learning metrics in correctly diagnosing and solving the learning crisis unfolding in developing countries (Pritchett, 2015; Figlio et al., 2016).

References

Ahsan, Md Nazmul, Rakesh Banerjee, and Siddharth Hari, "Social Promotion and Learning Outcomes: Evidence from India," 2018. Working paper.

Bansal, S. and S. Roy, "The tyranny of metrics: Why we learning nothing from the learning outcome data," *Financial Express, Feb 28*, 2019.

Bau, Natalie, Jishnu Das, and Andres Yi Chang, "New Evidence on Trajectories in a Low-Income Setting," 2021.

Bridgespan, "Changing the Public Discourse on School Learning: The Annual Status of Education Report," 2018.

Das, Jishnu, Tahir Andrabi, and Benjamin Daniels, "Human Capital Accumulation and Disasters: Evidence from the Pakistan Earthquake of 2005," 2020. Working paper.

Engzell, Per, Arun Frey, and Mark D Verhagen, "Learning loss due to school closures during the COVID-19 pandemic," *Proceedings of the National Academy of Sciences*, 2021, *118* (17), e2022376118.

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

Johnson, Doug and Andres Parrado, "Assessing the assessments: Taking stock of learning outcomes data in India," *International Journal of Educational Development*, 2021, *84*, 102409.

Kane, Thomas J and Douglas O Staiger, "The promise and pitfalls of using imprecise school accountability measures," *Journal of Economic perspectives*, 2002, *16* (4), 91–114.

_ **and** _ , "Volatility in school test scores: Implications for test-based accountability systems," *Brookings papers on education policy*, 2002, (5), 235–283.

Kingdon, Geeta Gandhi, "The private schooling phenomenon in India: A review," *The Journal of Development Studies*, 2020, pp. 1–23.

Muralidharan, Karthik, "The State and the Market in Education Provision: Evidence and the Way Ahead," 2019. Working paper.

_ **, Abhijeet Singh, and Alejandro J Ganimian**, "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 2019, *109* (4), 1426–60.

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

Pratham, *Annual Status of Education Report 2008 (Rural)* 2009. Pratham, New Delhi.

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

Pritchett, Lant, "Creating education systems coherent for learning outcomes: Making the transition from schooling to learning," *Research on Improving Systems of Education (RISE)*, 2015.

Reardon, Sean F, John P Papay, Tara Kilbride, Katherine O Strunk, Joshua Cowen, Lily An, and Kate Donohue, "Can Repeated Aggregate Cross-Sectional Data Be Used to Measure Average Student Learning Rates? A Validation Study of Learning Rate Measures in the Stanford Education Data Archive. CEPA Working Paper No. 19-08.," *Stanford Center for Education Policy Analysis*, 2019.

Rossiter, Jack, Martin Woodhead, Caine Rolleston, and Rhiannon Moore, "Delivering on Every Child's Right to Basic Skills, Summative Report," 2018. Young Lives, Oxford.

Shah, Manisha and Bryce Steinberg, "The right to education act: Trends in enrollment, test scores, and school quality," *AEA Papers and Proceedings*, 2019, *109*, 232–38.

Singh, Abhijeet, "Myths of Official Measurement: Auditing and Improving Administrative Data in Developing Countries," 2020. Working paper.

UIS, "Disruption Due to School Closure," 2022. https://covid19.uis.unesco.org/disruptions-due-to-school-closures/.

World Bank, *Wold development report 2018: Learning to realize education's promise* 2018. The World Bank, Washington DC https://www.worldbank.org/en/publication/wdr2018.

Appendix

Figures

Figure A1: ASER assessment tools

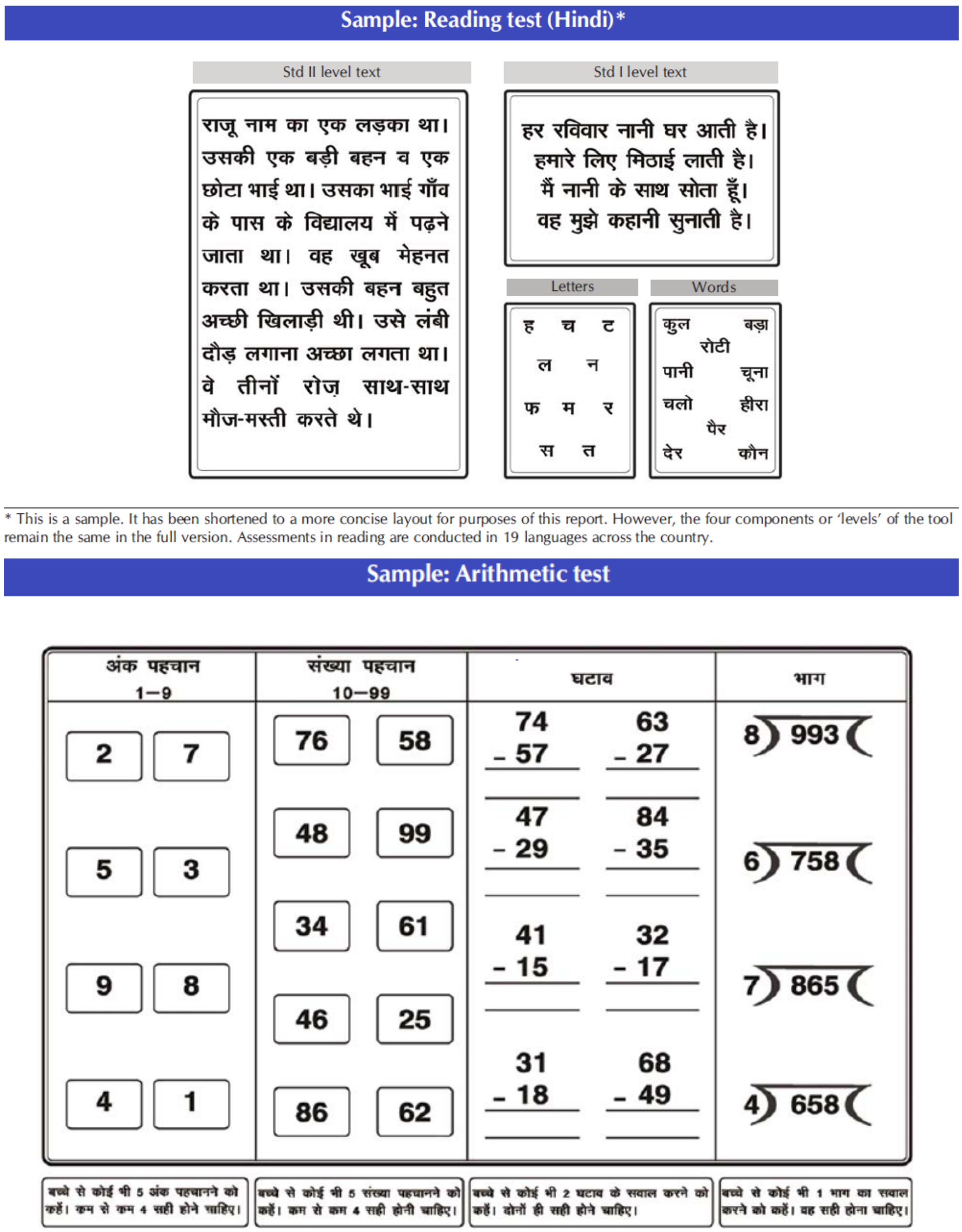
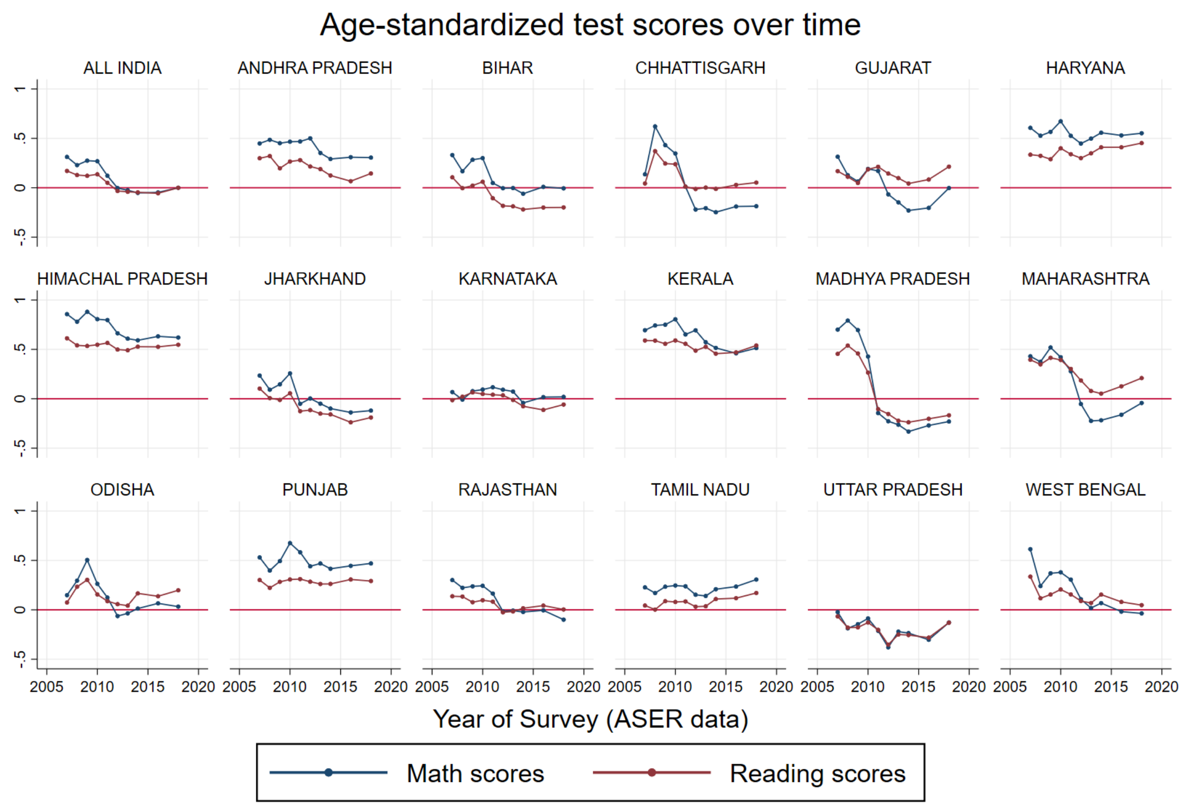
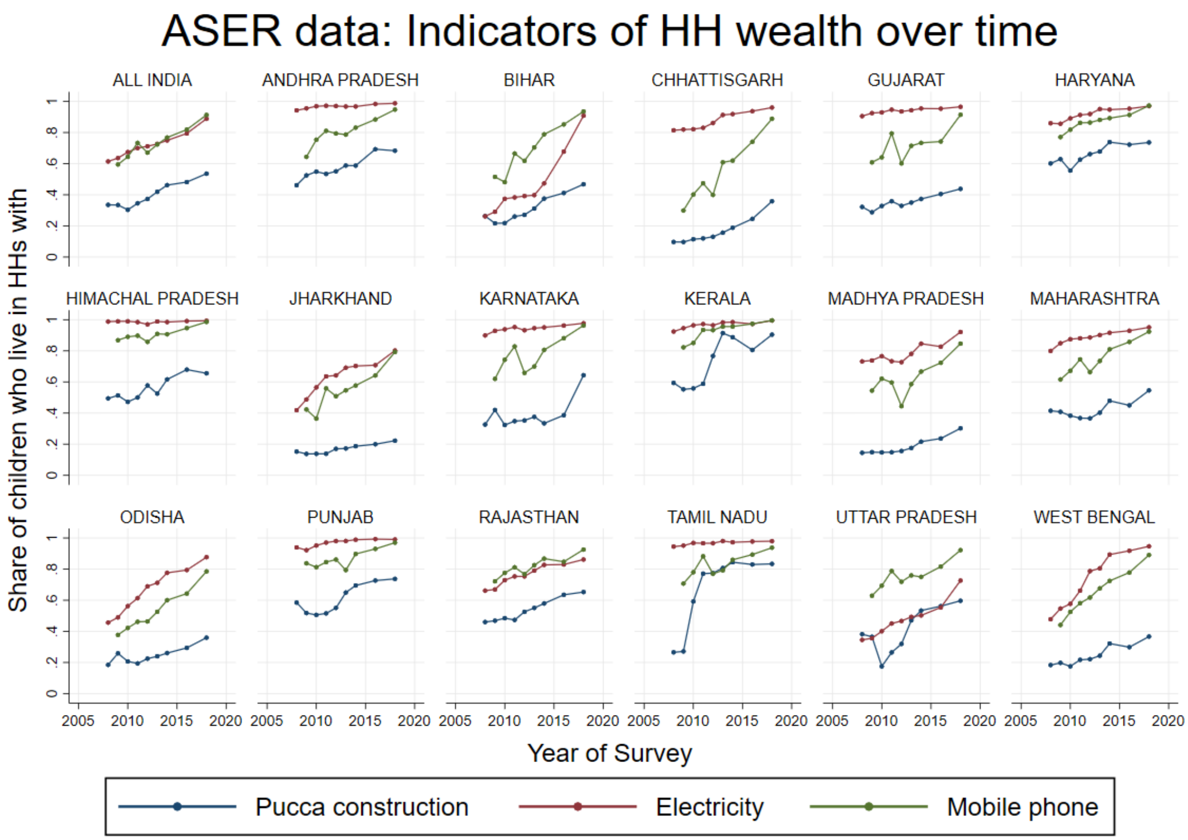


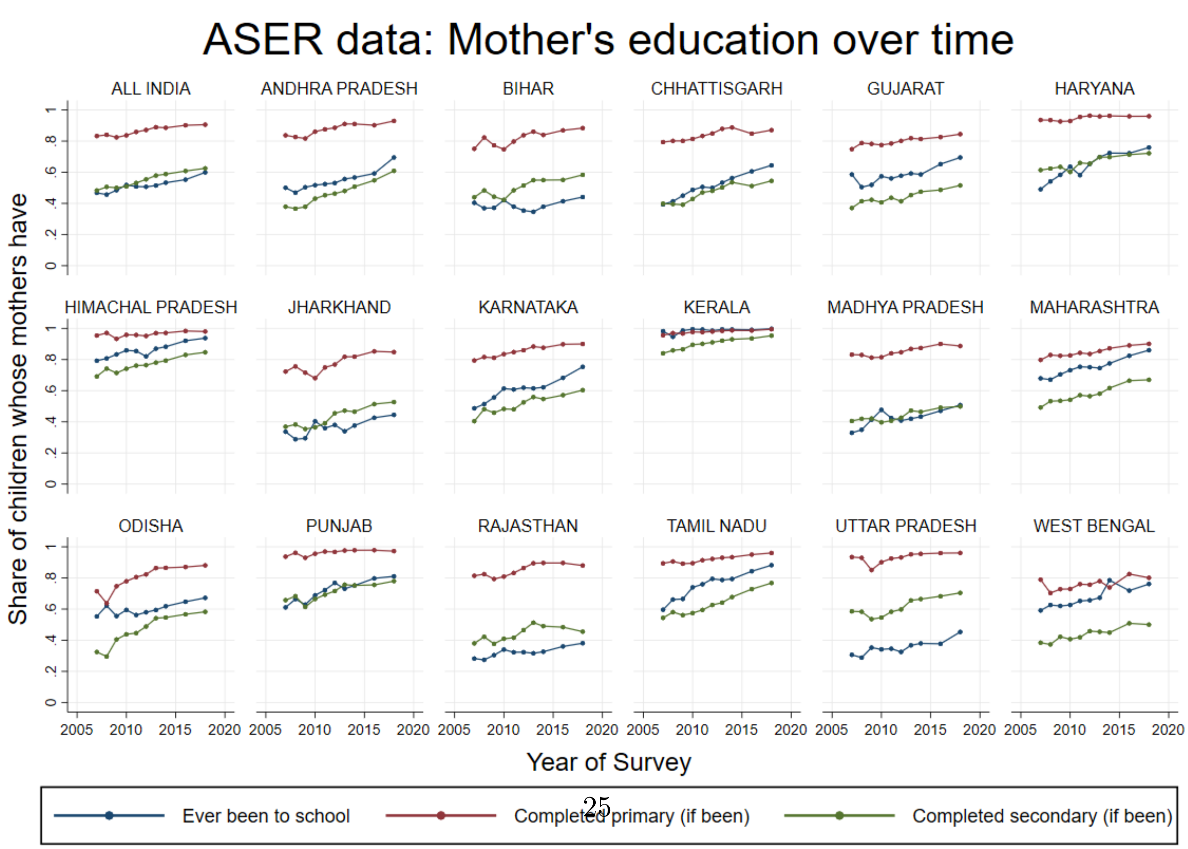
Figure A2: State-wise graphs of learning levels over time



Note: Scores are in s.d. units wrt the national distribution in 2018

Figure A3: HH characteristics over time





A Sources of learning outcome data

Besides ASER and the IHDS, there exists a third source of nationally-representative learning assessment data in the country - the government-run National Achievement Survey (NAS). Conducted roughly every three years since 2001 by the National Council for Educational Research and Training (NCERT), the NAS tests children studying in government schools on grade-appropriate learning outcomes. Table A1 presents a comparative summary of the ASER, IHDS and NAS data sets.

Table A1: Summary of nationally representative learning outcome surveys in India

	NAS	**ASER**	**IHDS**
Survey years	Roughly every 3 years since 2001. Only cycle 3 (2009-2013) and cycle 4 (2014-2016) are directly comparable.	2005-2014, 2016 and 2018	2006 and 2012
Approximate sample size	2,200,000	320,000	11,693
Population for which results are representative	Students attending govt and govt-aided schools in classes 3, 5, and 8	All children ages 5-16 in rural districts	All children ages 8-11
Learning outcomes data collected	Grade-appropriate math and language competency; self-administered paper and pencil test	Basic literacy and numeracy; administered orally and one-on-one	Identical to ASER
Other data collected	School infrastructure	School enrolment, basic household assets and parents' education	Rich set of household information such as employment, expenditure, etc.
Field staff	State education officials and teacher candidates	Partner organizations and volunteers	Full-time trained survey team

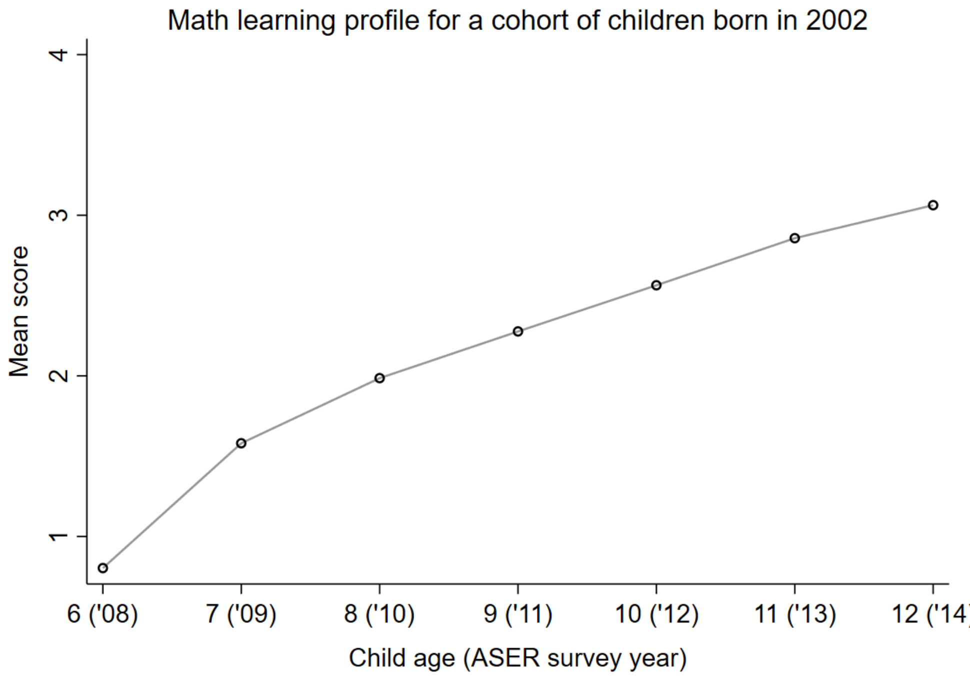
I do not use data from the NAS in this paper for three reasons. First, the NAS data series cannot be used to study trends in learning levels due to changes in scope, methodology, and reporting over time. The only directly comparable NAS survey rounds are cycle 3 (conducted between 2009-13 across different grades) and cycle 4 (conducted between 2014-16 across different grades). However, student-level data from the NAS is not in the public domain, and standardized measures of within-NAS changes in learning levels cannot be constructed using aggregate statistics presented in NAS reports. Second, since the NAS employs a different sampling strategy and learning assessment tool than the ASER/IHDS surveys, estimates are not directly comparable across these surveys. Finally, as shown by Johnson and Parrado (2021), data quality of NAS assessments appears to poor. In particular, estimates of learning levels in the NAS appear to be artificially high, and NAS state rankings display almost no correlation with state rankings based on ASER, IHDS, or net state domestic product per capita.

B What do learning profiles look like in ASER data? The case of Madhya Pradesh

Focusing on how learning levels of fixed birth cohorts of children evolved as they progress through school can tell us something about the nature of the learning decline observed in ASER data before 2012. A learning profile tracks test scores of a given child or a stable group of children as they get older and go from one grade to the next.[21] Figure A4 illustrates what a typical learning profile looks like in ASER data. For a cohort of children born in 2002, it plots average test scores between the ages of 6 (beginning of primary school) and 12 (end of primary school). The slope of the learning profile between any two consecutive ages, which measures the gain in test score for each additional year of schooling, is called the 'learning rate'. From figure A4, it is easy to see that the typical learning profile has non-negative learning rates across all adjacent ages. This is to be expected - with each additional year of schooling, children are usually able to score higher marks on tests of basic math and reading ability.

[21] ASER produces estimates of learning outcomes that are representative of all 5-16 year-old children at the district level in each year. Reardon et al. (2019) shows that cohort growth measures are generally good proxies for longitudinal growth measures, especially when student mobility rates are low.

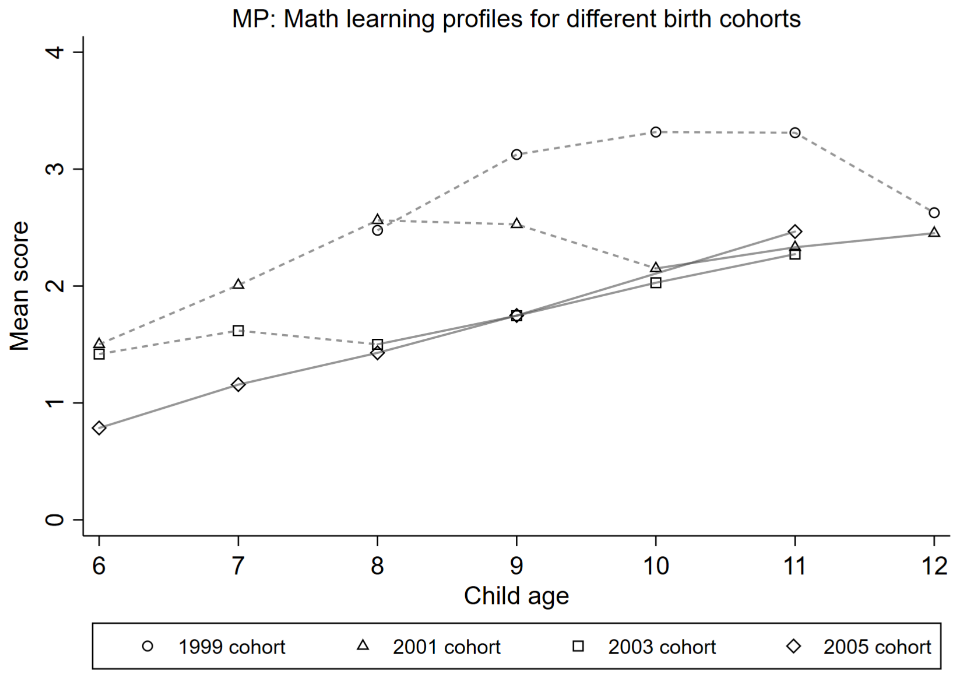
Figure A4: The typical learning profile



Notes: This figure plots average math scores for the cohort of children born in 2002 in the state of Tamil Nadu, tracked across the ages of 6 and 12 using data from ASER surveys conducted between 2008 and 2014 (survey years given in brackets).

Figure A5 plots multiple overlaid learning profiles for math scores for children in Madhya Pradesh across different different birth cohorts using survey data collected between 2007-2018. For each birth cohort, sections of learning profiles that use data from surveys conducted between 2007-2011 have been drawn using dotted lines, while sections that use data from survey conducted between 2012-2018 have been drawn as solid lines. The idea behind this exercise is to notice how the shape of learning profiles differs when using data from survey years that show large year-on-year declines (i.e. pre-2011) or from survey years during which learning levels were largely stable (i.e. post-2011).

Figure A5: Madhya Pradesh: Math learning profiles



Notes: This figures plots math learning profiles for different birth cohorts between ages 6 and 12 using ASER survey data collected between 2007-2018. Each point corresponds to a score-age pair for a given birth cohort. Within each learning profile, dashed line correspond to data collected between 2007-2011 and solid lines correspond to data collected between 2011-2018.

Begin by focusing on the learning profile for the 2005 birth cohort (diamond). The absence of any downward sloping segments in the learning profile shows that there were no instances of learning loss as children of this cohort progressed through primary school. Next, consider the learning profile for the 2001 birth cohort (triangle). Compared to children born in 2005, children born in 2001 entered school at age 6 with a much higher average level of math ability. However, this initial advantage was lost completely as they progressed through school, so that by age 10 there was no difference in average math scores between the 2001 and 2005 birth cohort. In particular, the 2001 cohort suffered learning losses between ages 8-10 which were so large that by the time they turned 10, they knew only as much as they did at age 7. Note that a similar pattern of decline can be observed at all ages between the 2009-2011 survey years across *all* birth cohorts.[22]

While true learning loss is certainly a possibility (Bau et al., 2021), the large magnitude and general nature of the decline is surprising given that (i) ASER's assessment tools test floor-level math and reading ability, and (ii) India has near-universal rates of primary school enrollment (so

[22] A similar pattern holds for reading scores, also across all birth cohorts.

it is likely that children were not out of school during this time). If this data reflects reality, the implication is that children across *every age group*[23] in MP forgot how to do basic arithmetic over a short period of 1-2 years despite attending school regularly. I find this to be implausible.

Finally, consider the consistent pattern outlined by the solid (constructed using data collected between 2011-2018) and dashed (constructed using data collected between 2007-2011) portions of learning profiles. For each birth cohort, solid lines are always upwards sloping and seem to follow a common lower envelope. On the other hand, dashed portions start at higher levels before falling to meet the common lower envelope exactly at the time of the 2011 survey. This implies that learning profiles of children across all birth cohorts converge after 2011, irrespective of their age at the time of the 2011 survey. This pattern is consistent with the hypothesis that learning profiles constructed using pre-2011 data are biased upwards, but eventually converge downwards to the 'real' learning profile (given by the common lower envelope) which are constructed using higher quality data collected after 2011.

C Constructing a measure of 'noise' in test scores changes

This section presents an analysis of how a measure of non-survey measurement error in year-on-year changes in the district-level test scores in ASER data has evolved between 2007-2018 using techniques outlined in Kane and Staiger (2002a). It supplements the analysis of non-survey measurement error in estimates of test scores levels in ASER data presented in the main body of the paper.

Assume that average test scores for a district (or state) at time t, y_t, consist of a fixed component that does not change over time, α; a persistent component v_t; and a purely transitory component ϵ_t. The persistent component, v_t, starts where it left off in the previous year, but is subject to a new mean-zero innovation each year, given by u_t. The purely transitory component, ϵ_t, is i.i.d. Thus, the average district-level test score in a given year equals:

$$y_t = \alpha + v_t + \epsilon_t; \quad v_t = v_{t-1} + u_t \tag{5}$$

The total variance of test score is equal to the sum of a persistent variance term (given by σ_u^2) and a transitory variance term (given by σ_ϵ^2).

[23] This is true even for children at age 6. Given that children this young would not have been exposed to changes in the schooling system imposed by the RTE, it is unclear the cohort of children who turned 6 in 2011 have lower average learning levels compared to all other birth cohorts that came before them.

$$var(y_t) = \sigma_y^2 = \sigma_u^2 + \sigma_\epsilon^2 \tag{6}$$

If we assume that changes more than one year apart are uncorrelated, we can calculate the total variance of changes in test scores as:

$$var(y_t - y_{t-1}) = var(\Delta y_t) = \sigma_u^2 + 2\sigma_\epsilon^2, \tag{7}$$

and the covariance of changes in test scores one year apart as:

$$cov(\Delta y_t, \Delta y_{t-1}) = cov(u_t + \epsilon_t - \epsilon_{t-1}, u_{t-1} + \epsilon_{t-1} - \epsilon_{t-2}) = -\sigma_\epsilon^2 \tag{8}$$

Equations 7 and equation 8 can then be used to estimate the share of the transitory component in the total variance in changes by the following formula:

$$-2 * Corr(\Delta y_t, \Delta y_{t-1}) = -2 * \frac{Cov(\Delta y_t, \Delta y_{t-1})}{var(\Delta y_t)} = \frac{2\sigma_\epsilon^2}{\sigma_u^2 + 2\sigma_\epsilon^2} \tag{9}$$

That is, given an estimate of the correlation in changes in test scores in two consecutive years, we can estimate the proportion of the variance in estimates of changes that is due to non-persistent factors (i.e. the share of 'noise') by multiplying that correlation by –2. Table A2 lists the average share of 'noise' in estimates of district-level changes in math and reading scores in ASER data across two time periods - between 2007 and 2011 and between 2012-2018.

Table A2: Share of noise in estimates of district-level changes in math and reading scores

	2007-2011	2012-2018	Overall
Average share of 'noise' in changes in math scores	69.7%	45.2%	58.8%
Average share of 'noise' in changes in reading scores	69.2%	61.7%	65.6%

Estimates from table A2 show that estimates of changes in district-level age-standardized math and reading scores in ASER data are rather noisily estimated in general. Between 2007 and 2018, the share of 'noise' was 58.8% in the estimates of changes in math scores and 65.6% for changes in reading scores. However, the share of 'noise' was significantly lower in estimates constructed using data from surveys conducted between 2012-2018 when compared to surveys conducted between 2007-2011. Across these two time periods, the average share of 'noise' reduced from 69.7% to only 45.2% for estimates of changes in math scores and from 69.2% to 61.7% for estimates of changes in reading scores.

Note that this method of constructing a measure of non-sampling measurement error in changes in test scores relies on somewhat stronger assumptions than the more general analysis of measurement error in levels of test scores (presented in the main body of the paper; also due to Kane and Staiger, 2002a). In particular, the present analysis relies on the assumption that changes in test scores that are more than one year apart are uncorrelated. This amounts to assuming that the the u_t and ϵ_t terms are not serially correlated. The presence of positive auto-correlation in either u_t or ϵ_t terms would induce downward bias in our estimate of the proportion of variance due to transitory shocks.