Assessing the Assessments: Taking Stock of India's Learning Outcomes Data

Doug Johnson and Andrés Parrado

Overview

- In India, there are two main sources of data on learning outcomes: ASER and NAS
- We assess the reliability of ASER and NAS
- We first compare ASER and NAS to each other (and IHDS)
- We then decompose variance in changes in ASER averages
- We find that:
  - NAS scores appear unrealistically high and contain little information about relative state performance
  - ASER scores are reliable measures of learning outcomes but a bit noisier than one would expect based on sample size

Basics of ASER, NAS, and IHDS

Comparing the datasets to each other

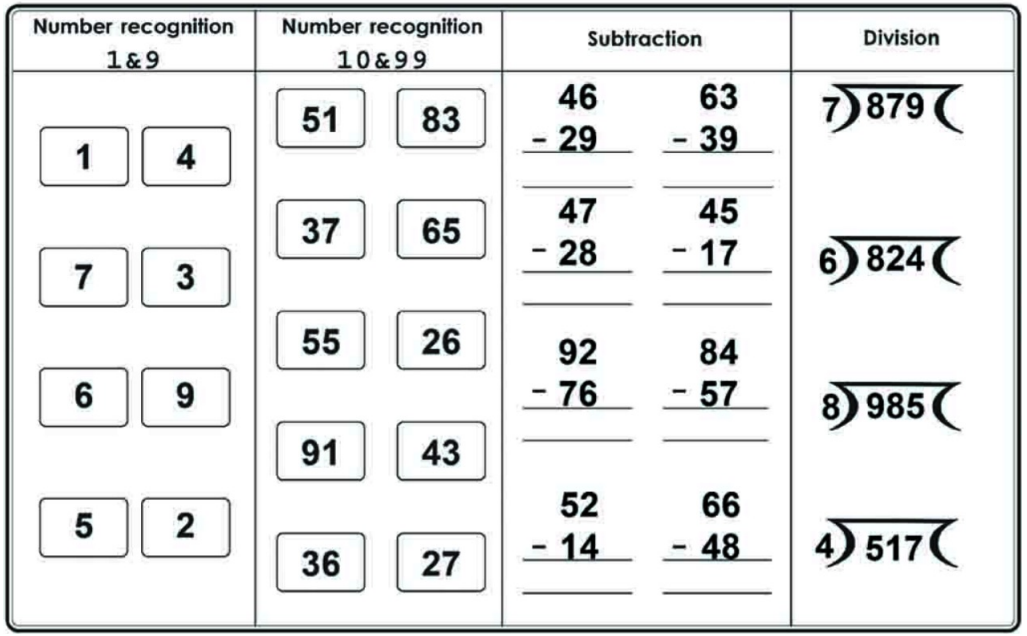
Assessing ASER reliability

Policy implications

ASER

- Implementation:
  - ASER Centre with the help of partner organizations (often DIETs) and volunteer surveyors
- Assessment tool:
  - Oral
  - Measures basic literacy and numeracy
- Sampling:
  - Only in rural areas
  - Villages randomly selected
  - In each village, households randomly selected using right-hand rule
  - All children 5-16 in selected hhs assessed
  - ~3.2 lakh children tested each year
- Frequency:
  - Every other year

Sample ASER math assessment



Number recognition 1&9	Number recognition 10&99	Subtraction	Division
1 4	51 83	46 − 29 ; 63 − 39	7)879
7 3	37 65	47 − 28 ; 45 − 17	6)824
6 9	55 26	92 − 76 ; 84 − 57	8)985
5 2	91 43	52 − 14 ; 66 − 48	4)517
	36 27		

National Achievement Survey (NAS)

- Implementation:
  - Government-run (with DIET students)
- Assessment tool:
  - Paper and pencil
  - Questions not publicly released, but seek to measure grade-level competency
- Sampling:
  - Government and private aided schools randomly selected using UDISE
  - In each school, up to 30 students randomly selected in grades 3, 5, and 8
  - ~ 22 lakh students assessed
- Frequency:
  - Conducted in its current format only in 2017

India Human Development Survey (IHDS)

- Implementation:
  - Independent, using paid surveyors
- Assessment tool:
  - Same as ASER
  - In addition, a variety of other household info collected
- Sampling:
  - Random selection of village / wards
  - Within each village, household randomly selected using household-listing
  - All children 8-11 assessed
  - ~12k children assessed
- Frequency:
  - 2011-12

Basics of ASER, NAS, and IHDS

Comparing the datasets to each other

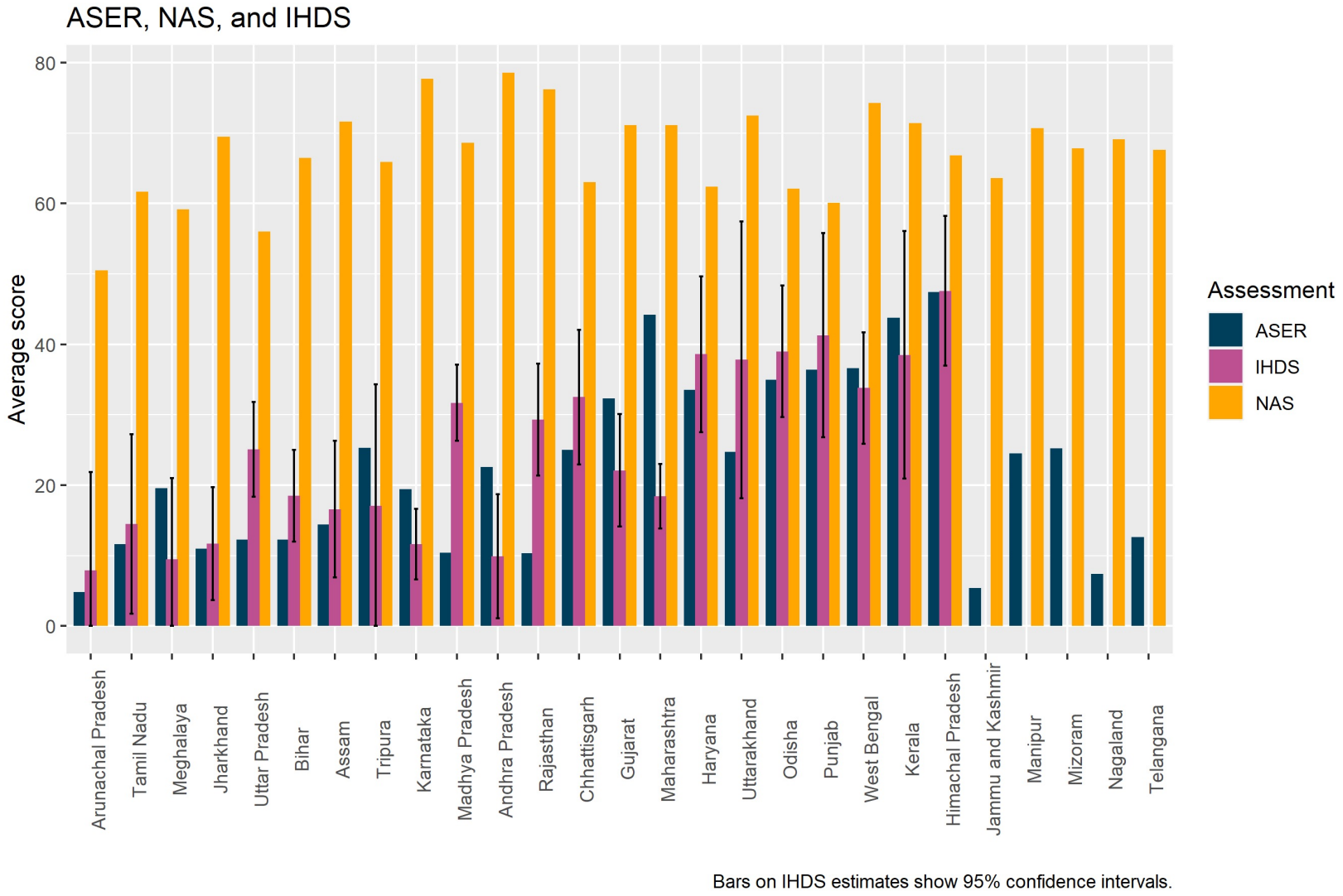
Assessing ASER reliability

Policy implications

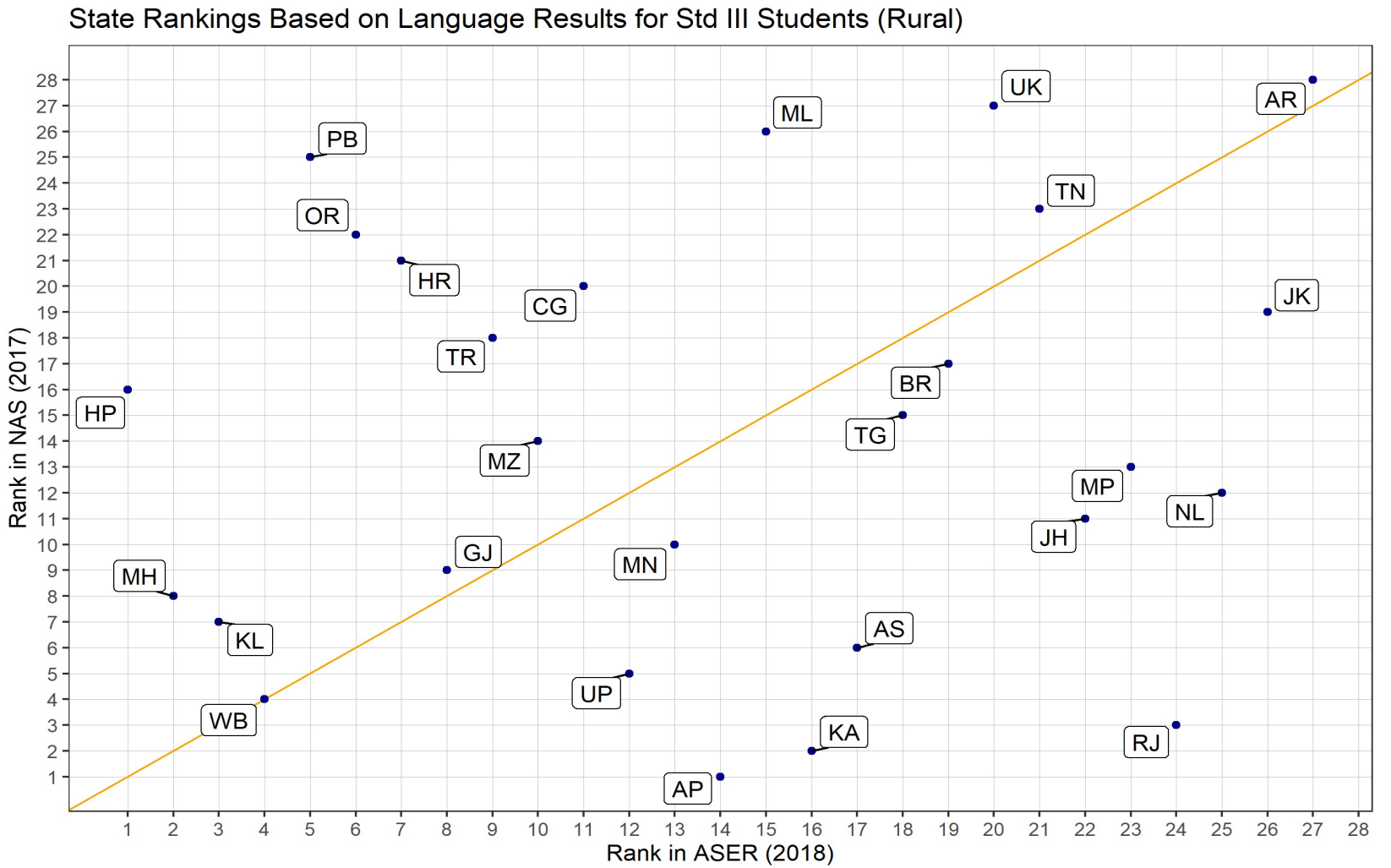
Making the datasets as similar as possible

- To make the samples as similar as possible we restrict the samples to only include:
  - **Grade 3 reading outcomes** because achieving the highest level of ASER corresponds to a 2nd grade reading level
  - **Rural areas** because ASER does not include urban areas
  - **Government (and private aided) schools** because NAS excludes private schools
- Despite these restrictions, there may still be differences in:
  - What is tested
  - Which students results are representative of (due to attendance)

ASER and IHDS are very similar but NAS scores are much higher



NAS scores and rankings display almost no correlation with ASER (or IHDS)



Why are ASER and NAS so different?

Sampling error?	✖ Sample sizes are huge
ASER non-sampling error?	✖ ASER and IHDS highly correlated
Differences in latent trait being measured?	? Possibly, but ASER reading and math highly correlated
NAS non-sampling error?	*Most likely*

Basics of ASER, NAS, and IHDS

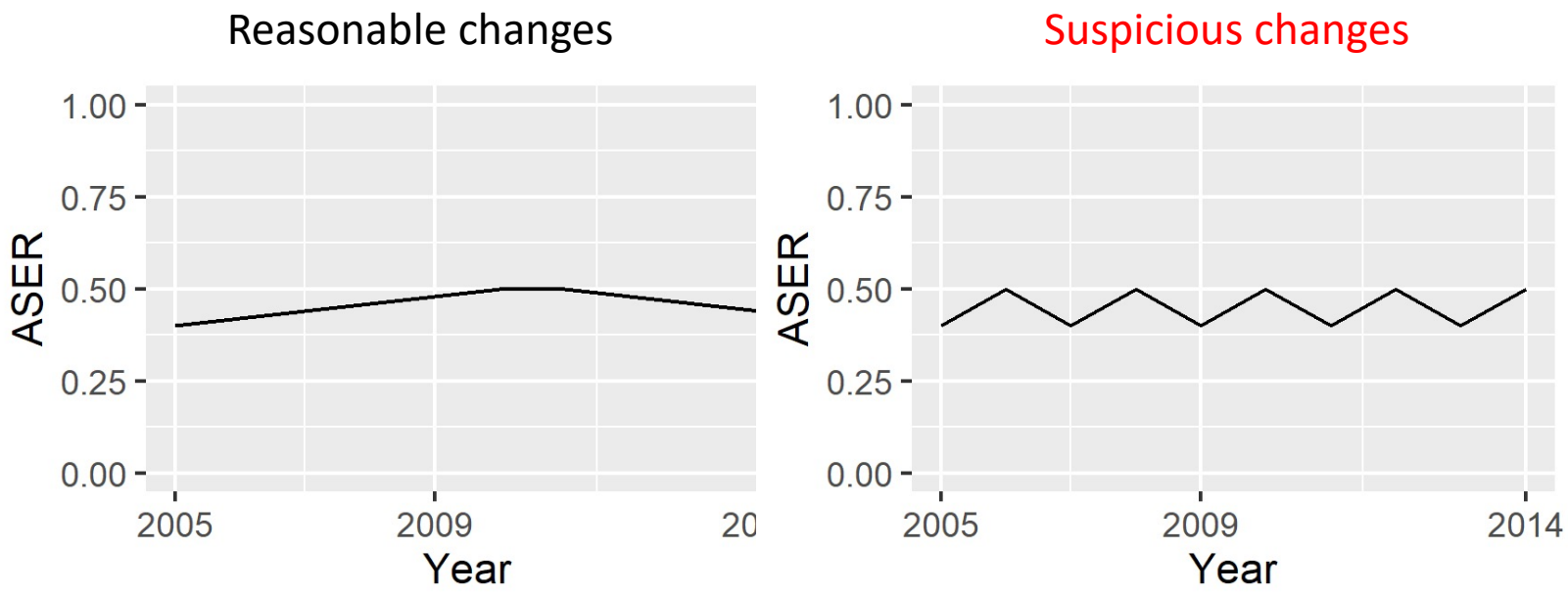
Comparing the datasets to each other

Assessing ASER reliability

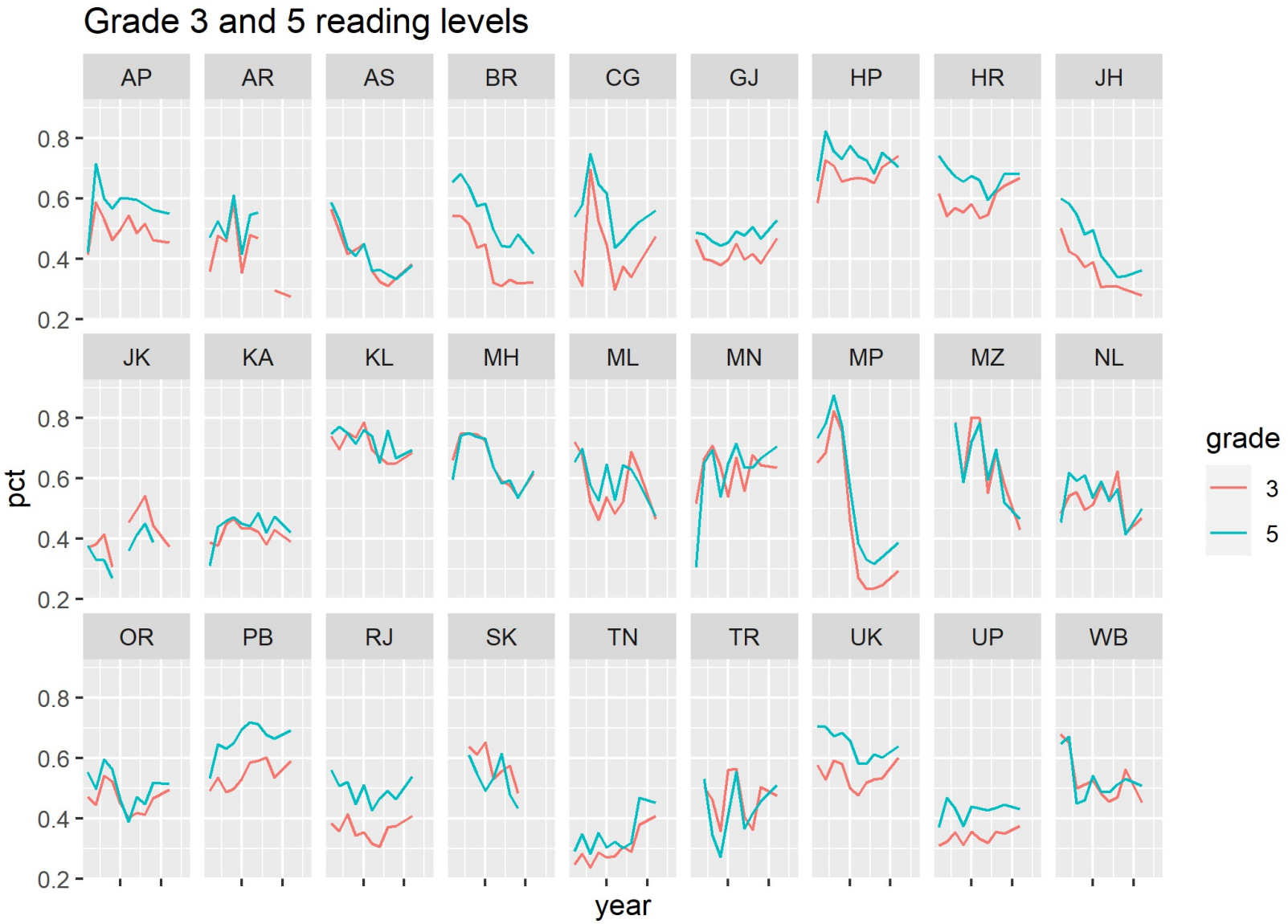
Policy implications

Analyzing ASER's internal reliability

- We don't have another multi-year dataset to compare ASER to but we can look at ASER data over time
- If year to year changes are often immediately reversed, we might suspect the "changes" are actually measurement noise

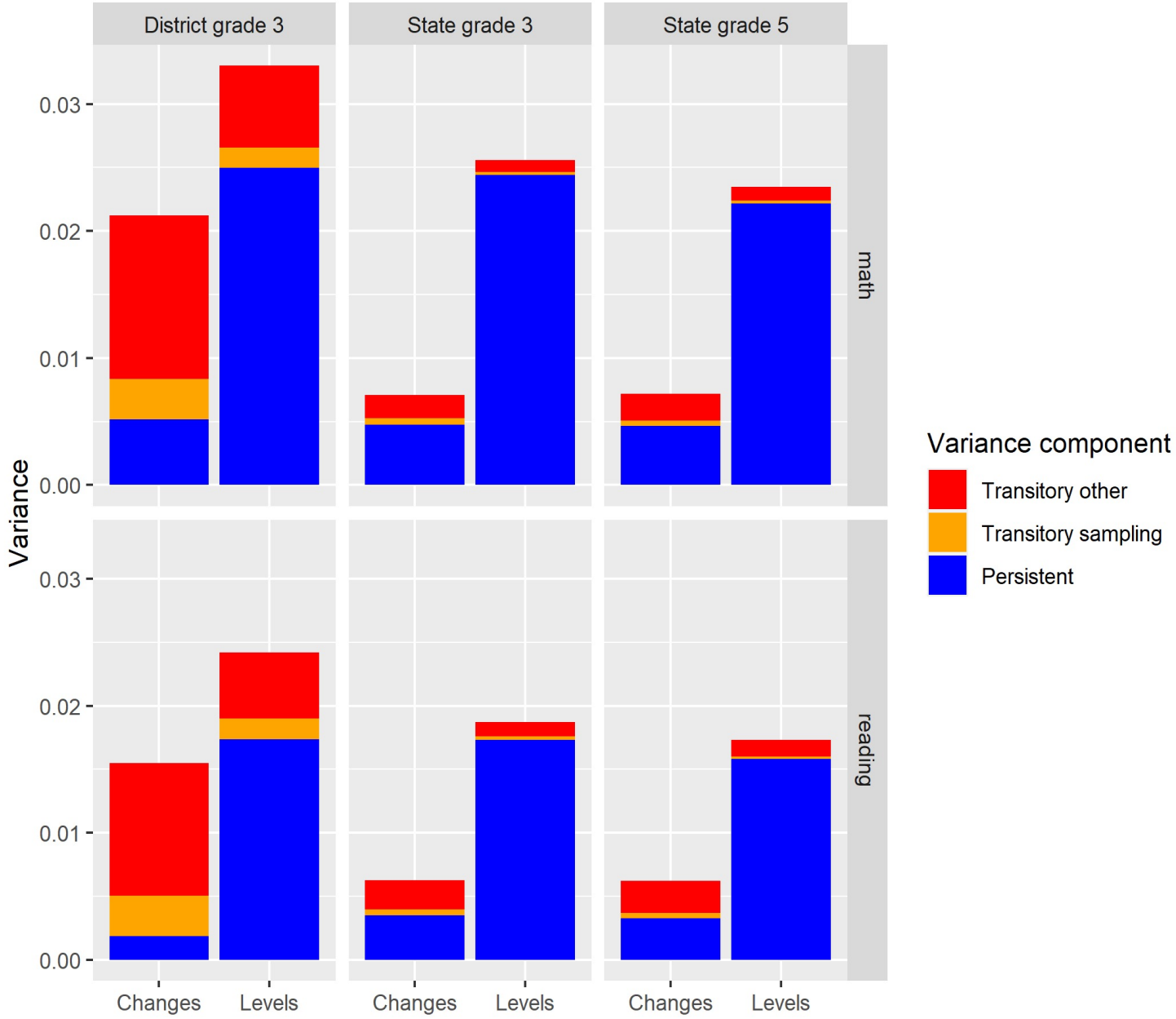


ASER trends over time



Quantifying “persistence”

- To quantify the share of suspicious and reasonable changes we use two methods from Kane and Staiger (2002)
- Both methods decompose the variance in changes in scores into “persistent” and “transitory” components
- We argue that transitory changes are likely due to measurement error
  - Most education policies are for multiple years
  - Differences between cohorts explain a very small portion of changes


District grade 3
State grade 3
State grade 5
math
reading
Variance
0.03
0.02
0.01
0.00
Changes
Levels
Variance component
Transitory other
Transitory sampling
Persistent

Basics of ASER, NAS, and IHDS

Comparing the datasets to each other

Variance decomposition of ASER

Policy implications

Implications for use of these datasets and future data collection

- Using these datasets
  - NAS – Exercise extreme caution when using this dataset!!
  - ASER – Be cautious when comparing changes over time (or using district data)
- Potential future data collection
  - Non-sampling error >> sampling error. Theoretically, a survey with 0 non-sampling error could achieve higher precision with a fraction of sample size.
  - Unless source of noise is identified and corrected, future rounds unlikely to yield useful data

Thanks!

Kane and Staiger Method 1

- Assume scores made up of a fixed effect (α), a "persistent" change component (v_t) and a "transitory" change component (ε_t):
  - $y_t = \alpha + v_t + \varepsilon_t$
- Assume persistent component follows random walk:
  - $v_t = v_{t-1} + u_t$
- Then the share of total variance due to transitory variance in changes is…
  - $\frac{Var(\Delta\varepsilon)}{Var(\Delta y)} = \frac{2\sigma_\varepsilon^2}{\sigma_u^2+2\sigma_\varepsilon^2} = -2 * \mathrm{corr}(\Delta y_t, \Delta y_{t-1})$

Kane and Staiger Method 2

- Assume correlation between current year scores and previous year scores, ρ_1, reflects transitory changes + persistent changes
- ..But decay in autocorrelation after one lag (i.e. difference between ρ_1 and ρ_2) reflects true changes
- Then:
  - $\sigma^2_{pers} = \frac{\sigma^2_y * \rho^2_1}{\rho_2}$