District level variation in learning

To what extent does the changing age distribution affect learning levels?

Shannon Stackhouse Flores, ASER research fellow

September 16, 2009

In the 2008 ASER report, Dr. Wilima Wadhwa commented on a problem in the learning assessment data: the high degree of variation in learning levels within many districts between 2006 and 2007, and then again between 2007 and 2008. Among the possible causes of this "jumpiness" in the data, one major hypothesis was that the large observed variation in age distribution within districts might be contributing.

The household-based design of the survey, as well as budget and time constraints have an impact on total sample size, as well as the ability to control the population of children included in the sample. Dr. Wadhwa notes that in an ideal environment, surveyors would be able to develop a complete listing of households within each village and district, then determine the households to be surveyed to ensure the desired distribution of children by age and/or standard. Given the importance of quick turnaround and rapid distribution in achieving the goals of ASER, the creation of a full house list is not possible. Random selection of households occurs at the same time as the survey itself, by the volunteers themselves, as described throughout the ASER reports. Thus in the same district, distribution of children by age and class might vary dramatically from year to year.

This paper seeks to address the concern that changing proportions of children by age from year to year may have affected learning levels, contributing to the great fluctuation in learning levels observed in some districts.

Table 1
Age Fluctuation: % by Age Standard 1. Sample district.

	2006	2007	2008
5	15	29	19
6	72	44	36
7	5	16	27
8	5	8	9
9	2	2	4
10	1	1	2
11	0	0	2
12	0	0	1
13	1	1	0
14	0	0	0
15	0	0	0
16	0	0	0

Source: Wadhwa, W. challenges of generating district level estimates, ASER Report 2008

Figure 1. Age Fluctuation: % by Age Standard 1

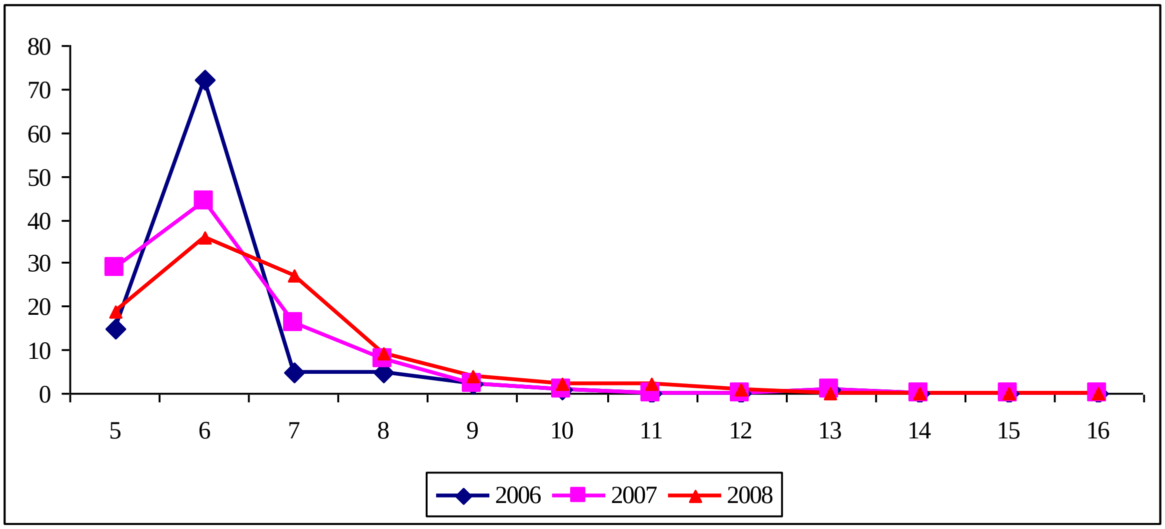


Table 2. Example of extreme district-level change in percent of children performing at basic learning levels, 2006 through 2008			
Standard 1	**2006**	**2007**	**2008**
% of children who can identify letters	73%	76%	29%
% of children who can identify numbers 1-9	-	70%	27%
Standard 5			
% of children who can read paragraphs	76%	68%	37%
% of children who can subtract	72%	62%	22%

Focusing on the 2007 and 2008 data, in order to quantify the extent to which the change in learning level was affected by change in the age distribution, a simple regression model was constructed with district aggregate learning levels as the outcome (dependent) variable. The analysis was run separately for Standards 1 & 2 and Standards 3-5, to correspond with the data presented in the annual report. For each district[1], the percentage of children at each age able to perform at the specified learning level was calculated. For Standards 1 & 2, the outcome variables were: percentage able to identify letters or more for reading, and percentage able to identify numbers 1-9 for math. For Standards 3-5, the outcome variables were percentage able to read paragraphs and percentage able to perform two-digit subtraction.

The model specifically tested how much the change in the percentage from 2007 to 2008, per district, was attributable to the change in percentage of each age group[2]. In the simplest model, only the

[1] The raw variable "district" was first converted to a stratum variable, combining the state and district codes into a unique identifier.

[2] A linear regression was used. Example: Δ(% children able to read at least letters) = β0 + Δ(% children in Standards 1 and 2 who were 5 years old) β1 + Δ(% 6-year olds) β2 … + Δ(% 12-year olds) β8 + ε. Children aged 13 to 15 were included in regressions for standards 3 – 5.

age variables were used as independent variables. In a slightly more complicated model, run for Standards 1 & 2, other variables were included: percentage of children whose mothers had attended school; number of children in the family; and the percentage of children that were in Standard 2 rather than Standard 1. These variables were chosen because of the available controls in the individual data sets, these had the greatest significance in trial regressions. Mother's education and number of children are known in development literature to effect schooling outcomes; including the percentage in Standard 1was a way of testing whether the changing class distribution might have had an effect independent of age. Other household characteristics were not aggregated, as they were only available in 2008. The models using these additional control variables performed no better than without, so the results reported here are based solely on the age-only regressions.

Table 3. Standards one and two; change in percentage of children who could read at least letters[3]; Linear Regression, B and (Standard errors)

	All districts	BH	KAR	MP	MH	OR	RJ	TN	UP
R2	0.011	-0.123	-0.274	-0.287	0.143	-0.242	-0.025	0.011	0.098
F	1.753	0.535	0.301	0.561	1.666	0.294	0.904	1.038	1.922
Δ Age 5	-0.885**	1.052	-1.089	2.693	-2.932*	0.147	-0.188	-8.361	-2.910
	(0.324)	(2.978)	(3.020)	(2.320)	(1.336)	(1.211)	(3.867)	(4.461)	(2.227)
Δ Age 6	-0.747**	1.017	-1.655	2.530	-1.609	0.316	0.592	-7.731	-3.300
	(0.317)	(3.070)	(2.974)	(2.243)	(1.312)	(1.046)	(3.703)	(4.372)	(2.231)
Δ Age 7	-0.735*	1.420	-1.578	2.858	-2.001	0.484	0.778	-7.504	-2.620
	(0.318)	(2.926)	(2.995)	(2.299)	(1.329)	(1.053)	(3.905)	(4.317)	(2.197)
Δ Age 8	-0.743*	1.582	-1.335	2.574	-1.563	0.381	0.626	-8.979	-3.422
	(0.317)	(3.078)	(2.917)	(2.287)	(1.314)	(1.128)	(3.841)	(4.623)	(2.212)
Δ Age 9	-1.050**	-0.211	-2.699	3.202	-1.454	1.378	0.076	-10.35	-3.894
	(0.383)	(3.122)	(3.592)	(2.357)	(1.407)	(2.874)	(4.479)	(5.646)	(2.353)
Δ Age 10	-0.885*	1.823	-0.565	2.231	-0.611	-1.037	2.645	-6.375	-1.586
	(0.409)	(3.039)	(3.468)	(2.546)	(2.067)	(3.240)	(4.378)	(5.387)	(2.640)
Δ Age 11	0.176	2.735	0.745	3.269	-3.583	3.281	-0.225	- 6.261	-3.505
	(0.24)	(4.175)	(8.131)	(3.483)	(3.836)	(4.774)	(6.163)	(8.553)	(3.831)
Δ Age 12	-1.332	-0.678	-2.545	1.317	-1.227	-1.458	-4.97	4.223	-6.792**
	(0.682	(4.649)	(5.273)	(3.486)	(2.893)	(3.355)	(5.272)	(8.462)	(2.669)

Notes: * Significant at 5% level, ** significant at 1% level.

Adjusted R2 reported

[3] Generally speaking, R- squared and variable significance were even lower for Standards 1 and 2 Math; thus the results are not presented. Standards 3 and 5 math were more interesting than reading, so are presented here. For Standards 3-5 reading, Bihar showed R-squared = 0.267 but no ages were significant; Madhya Pradesh had R-squared equal to 0.199, and many variables were significant, but all were negative.

Table 4. Standards 3-5; change in percentage of children who could perform two digit subtraction ; Linear Regression, B and (Standard errors)

	All districts	BH	KAR	MP	MH	OR	RJ	TN	UP
R2	0.022	0.410	-0.035	0.267	0.071	-0.243	0.038	-0.202	-0.091
F	2.061*	3.149**	0. 919	2.460**	1.221	0.485	1.110	0.573	0.482
Δ Age 5	-0.081	-7.286	-29.16	-4.074	-10.771	-1.486	6.962	-8.229	2.935
	(11.535)	(3.987)	(32.38)	(5.628)	(10.039)	(8.411)	(5.225)	(17.552)	(5.689)
Δ Age 6	-1.328	-1.848	-10.57	-13.725**	-14.299	-5.870	5.731	-6.392	-2.324
	(11.551)	(3.515)	(24.47)	(4.752)	(9.955)	(7.046)	(5.189)	(11.646)	(4.212)
Δ Age 7	-2.079	-7.198	-6.436	-10.384*	-19.681*	-4.814	2.679	-11.070	-0.120
	(11.518)	(3.591)	(22.72)	(4.316)	(8.602)	(7.539)	(4.294)	(10.48)	(4.007)
Δ Age 8	-1.631	-5.325	-5.383	-9.101*	-16.804*	-4.982	3.418	-11.586	-0.005
	(11.541)	(3.497)	(23.16)	(4.222)	(7.929)	(7.284)	(4.584)	(10.777)	(4.052)
Δ Age 9	-1.298	-3.488	-4.512	-8.255	-17.29*	-5.659	3.810	-11.579	-0.450
	(11.538)	(3.332)	(23.49)	(4.152)	(8.133)	(7.142)	(4.603)	(10.784)	(4.051)
Δ Age 10	-1.461	-4.426	-4.634	-8.800*	-17.797*	-6.684	4.563	-11.678	-0.159
	(11.532)	(3.297)	(23.33)	(4.119)	(8.124)	(7.070)	(4.672)	(11.031)	(4.002)
Δ Age 11	- 1.713	-7.119	-4.565	-9.291*	-17.521	-6.514	4.976	-12.582	0.521
	(11.537)	(3.647)	(23.18)	(4.289)	(8.045)	(7.674)	(4.643)	(11.17)	(3.975)
Δ Age 12	-2.032	-8.484	-5.215	-8.711*	-15.063	-5.665	2.536	-14.850	-0.119
	(11.529)	(3.475)	(23.39)	(4.240)	(8.122)	(6.635)	(4.789)	(11.024)	(4.179)
Δ Age 13	-1.726	-0.204	-5.105	-6.060	-9.239	-4.732	2.272	3.460	-1.358
	(11.598)	(3.557)	(23.91)	(4.303)	(9.098)	(9.532)	(4.723)	(14.258)	(4.231)
Δ Age 14	-2.456	-8.445*	-8.560	-11.884*	-17.221	-3.763	3.349	-17.857	-0.078
	(11.631)	(4.439)	(24.16)	(4.826)	(8.832)	(9.229)	(7.135)	(13.871)	(4.846)
Δ Age 15	-0.731	-2.285	3.875	-3.509	-24.503	-1.858	8.831	- 7.357	0.754
	(11.435)	(4.824)	(33.20)	(6.476)	(11.488)	(12.260)	(7.394)	(20.80)	(5.431)

As shown above, for both groups, the model was run first for all districts, then separately for each of eight large states: Bihar, Karnataka, Madhya Pradesh, Maharashtra, Orissa, Rajasthan, Tamil Nadu, and Uttar Pradesh. Although the model was significant when run for all districts, R-squared was extremely low, and the model was significant for very few states. Overall, the pattern suggests that the difference in age is *not* systematically responsible for the fluctuation in learning between the two years. For Standards 3 – 5, the relationship between age and math performance did appear to have some significance in Bihar[4] and Madhya Pradesh, but even within these states, the pattern was not clear; i.e.

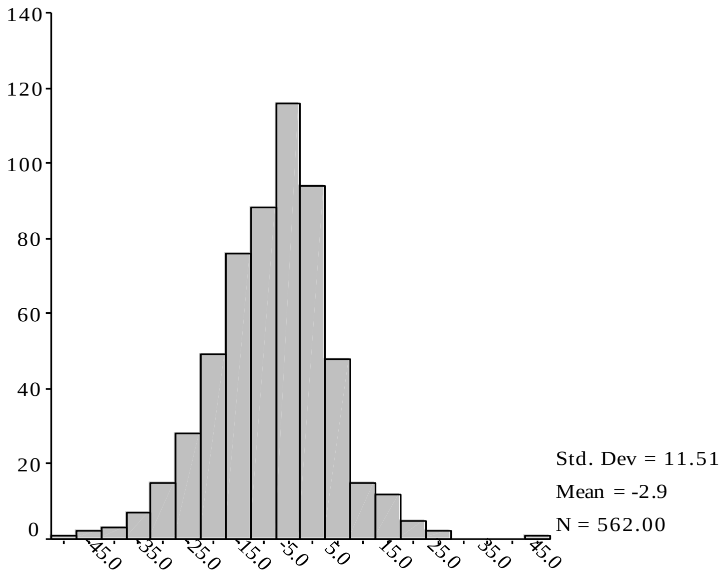
[4] In Bihar, the model was significant at the 10% level.

there is no evidence that having a higher proportion of younger students is systematically correlated with lower learning levels, or vice versa[5].

Within-District Differences in Learning

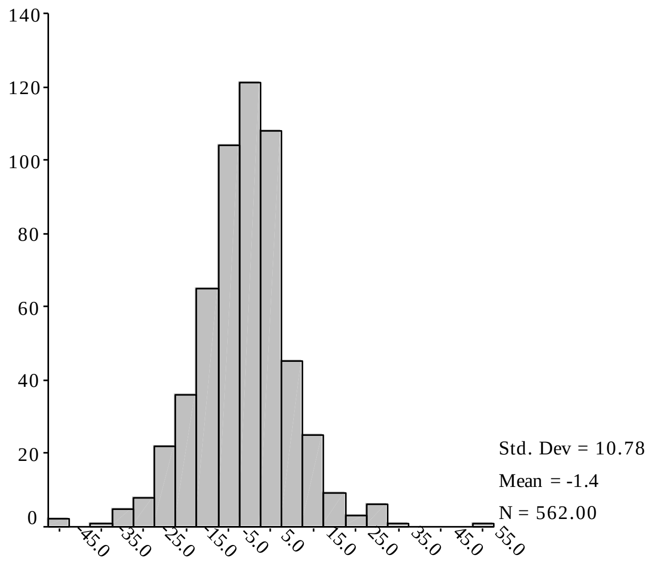
Again using Standards 1 and 2 as an example, when looking at the graph of change per district between 2007 and 2008 (below), it is heartening to note that the greatest number of districts experienced near 0% change in the percentage that could identify letters and the percentage that could identify numbers. On average, 3% fewer children per district could identify letters in 2008 than in 2007; and 1% fewer could identify numbers. The standard deviations are rather high, however, indicating that there are quite a few districts with higher changes. At the extremes, there were districts showing 30 to 40% increases, as well as 30 to 40% decreases for both reading and math.

Figure 2. Change across districts in percentage of children who could identify letters, 2007- 2008.



[5] In many states, the negative and positive coefficients on the use variables appeared fairly random. In some states, variables are either all negative or all positive.

Figure 3. Change across districts in percentage of children who could identify numbers, 2007- 2008.



To determine whether age was a factor in those districts that did vary greatly, the regression was run restricted to only those districts that showed greater than 20% change[6] in either direction. The regressions revealed no age effect.

Other Factors to Consider

When the same aggression was run separately for the 2007 aggregate district data and the 2008 aggregate district data, age did appear as a significant factor affecting learning, even in the presence of the controls. This suggests that while age is a factor within years, between years other factors contribute to more of the change than does age alone. It is not possible to check for many other factors, because they're not comparable across the years. It might be possible to explore this more between 2008 and 2009.

One possibility is that a variation in testing tools, sampling of children that is not due to age or standard, or differences in the way that volunteers assess learning. It was apparent in looking at the data, that districts with large changes in reading also showed large differences in math. When percentage change in children identifying numbers 1-9 was added to the regression for children identifying letters (test was done for Uttar Pradesh), the model fit increased dramatically, and math ability was the only significant variable.

[6] Roughly two standard deviations.

Table 5. Standards one and two; change in percentage of children who could read at least letters; Linear Regression, B and (Standard errors)

	Uttar Pradesh
R2	0.780
F	222.481**
Δ Age 5	-0.159
Δ Age 6	-0.064
Δ Age 7	-0.161
Δ Age 8	-0.080
Δ Age 9	-0.253
Δ Age 10	-0.242
Δ Age 11	0.179
Δ Age 12	-0.251
Δ % rec. numbers 1-9	0.940**

Academic achievements in reading and math are generally expected to be correlated for any individual child, so in and of itself a high correlation between these two variables does not indicate a cause for alarm. The fact that R- who squared is so high even at the district aggregate level, however, could indicate some hard to detect a difference in the testing instruments between years, the particular villages chosen, or (possibly more likely) that much of the variation is due to differences in how surveyors select children to test, or how they themselves assess learning[7]. This is especially likely since the same surveyor assesses both reading and math for each child. If one year of surveyor is more lenient than the surveyor assigned the district for the next year, the assessment of learning would change dramatically, but similarly for math and reading.

[7] The ASER staff has focused a lot of energy on standardizing the survey and assessment, through streamlined tools, rigorous trainings and documentation of the survey process. The quantification of surveyor bias, however, remains an important one to tackle.

Conclusion

From this analysis, we can be fairly certain that age alone is not contributing consistently to the fluctuation in learning levels within districts from year to year[8]. Although in some districts age might have contributed to differences in learning, on the whole it does not appear to have been a large problem. This allows the attention to shift toward identifying and quantifying other potential sources of bias as the survey continues to be refined or to move into its next stage.

It would perhaps be informative for the states with higher R-squared (Bihar and Madhya Pradesh) to go back and look more closely at the available information. It may be helpful to go back to the district aggregate and a regression in again with all possible variables or to create two groups: ages 5 – 10 and 11 – 15. This might lessen the extent to which the age variables are taking significance away from one another.

[8] Or at least from 2007 to 2008.