Evaluating the Reliability and Validity of the ASER Testing Tools

Shaher Banu Vagh
ASER Centre: www.asercentre.org

The Annual Status of Education Report (ASER), a nationwide survey of reading and math achievement of children from rural India has been conducted annually since 2005. ASER provides basic and critical information about rural Indian children's foundational reading skills and basic math ability. Given its scale and comprehensive coverage, it is a path breaking initiative as it is the only nationwide survey, albeit rural, which assesses the learning achievement of children in standards 1-8 (ages 5-16). The survey is conducted each year in the middle of the academic year (October to November) and the findings are made public, for most states, in the same school year (mid-January). The availability of results in the same school year is a tremendous feat for such a large survey, which enhances its potential as a tool to inform educational practice and policy.

The ASER test inference is about a child's level of foundational reading skills (akshar identification, word decoding, etc.) and basic math ability (number recognition, subtraction, and division). The content of the ASER-reading test, i.e. the selection of words, length of sentences and paragraphs, use of vocabulary is aligned to standard 1 and standard 2 level state textbooks and the ASER-math test is aligned to standard 1, 2, 3, and 4 level state textbooks. The tests are orally and individually administered and require 10 minutes of administration time. The tests are designed as criterion-referenced tests that categorize children on an ordinal scale indexing mastery in the basic skills of reading and number operations. The tests are designed to understand what students can do and the skills they have mastered. For instance the ASER-

Vagh, S.B. (2009). Evaluating the reliability and validity of the ASER testing tools. New Delhi: ASER Centre.
http://www.asercentre.org/sampling/precision/reliability/validity/p/180.html

reading test classifies children at the ‘nothing’, ‘letter’, ‘word’, ‘paragraph’ (grade 1 level text), and ‘story’ (grade 2 level text) level based on defined performance criteria or cut-off scores that allow examiners to classify children as masters or non-masters of any given level. For e.g. the inability to correctly identify 4 out of 5 akshars classifies the child at the ‘nothing’ level. The ASER math test classifies children at the ‘nothing’, ‘single digit recognition, ‘double digit recognition, ‘subtraction with carry over’, and ‘division’ level (see www.asercentre.org for testing tools and the annual reports for test administration details.).

The ASER testing tools have several advantages: they are simple, quick, cost-effective, and easy to train examiners to administer. All of these are desirable features (Wagner, 2003) as it makes feasible a survey of the scale and scope of the ASER (assessing about 700,000 children every year) and makes results available in a timely manner, which has the potential to inform educational practice and policy. However, several pertinent questions have been raised about the ASER testing tools in relation to their content and their statistical properties. Specifically, how robust are the ASER-reading and ASER-math testing tools? Do the ASER testing tools provide reliable and valid findings? The present report, therefore, aims to address these critical questions about the Hindi language and the math testing tools.

Defining Reliability and Validity:

The traditional notion of reliability is the *consistency* with which a test measures any given skill and thereby enables to *consistently* distinguish between individuals with regards to the ability or skill being measured. In other words, if it were possible and feasible to test children repeatedly using the same test, a reliable test would yield a *consistent* score across the repeated measurements. However, given that the ASER tests assess achievement of mastery rather than the relative standing of children in relation to their peers, reliability in this case is "the consistency of the decision-making process across repeated administrations of the test" (Swaminathan, Hambleton, & Algina, 1974). Hence, reliability in this case does not refer to 'test reliability' but rather to the 'reliability of decisions' (Huynh, 1976 as cited in Traub & Rowley, 1980) or 'decision consistency' (Swaminathan, Hambleton, & Algina, 1974) as the assessment here is about mastery or non-mastery of a level of reading or math.

Validity, on the other hand, indicates whether the test measures what it purports to measure, i.e. how well children's performance on a test support the conclusions we make about a specific ability or skill. For instance is the inference based on the ASER-reading test about children's *mastery or non-mastery of basic reading ability* valid? Is the inference based on the ASER-math test about children's *mastery or non-mastery of basic math ability* valid? Specifically, validity is an evaluation of a test inference and not of the test per se. A test can be put to different uses such as, examining average school performance or making diagnostic decisions about individual students. Each of these uses or inferences "has its own degree of validity, [and] one can never reach the simple conclusion that a particular test "is valid"" (Cronbach, 1971, p.447). Another way to think about reliability and validity is that when playing darts, consistently hitting the same spot, irrespective of position on the target board is akin to reliability and consistently hitting bull's eye or any other target of interest, is akin to validity. Reliability then is a necessary but not sufficient condition for validity.

Methods for Assessing Reliability and Validity

A traditional classical test theory approach to evaluating consistency or stability of performance across repeated assessments is termed test-retest reliability. Given that children are expected to be learning in schools we expect high stability coefficients over a shorter time span than a longer time span between two testing sessions. The test-retest reliability indexes the relative ranking of individuals across two testing occasions. However, it does not indicate the consistency of categorizing children at the different levels of mastery.

Given that the ASER-reading and -math tests are criterion-referenced tests the evaluation of reliability in this case, as noted earlier, is essentially a measure of the agreement between decisions made in repeated test administrations (Swaminathan, Hambleton, & Algina, 1974; Traub & Rowley, 1980). Swaminathan, Hambleton, & Algina (1974) suggested the use of Cohen’s Kappa to evaluate agreement between decisions across repeated test administration. The Cohen’s Kappa takes into account agreement due to chance and thus provides an estimate of consistency beyond chance (Cohen, 1960).

In addition, given that examiners categorize children at different levels of mastery an evaluation of inter-rater reliability, which is a measure of the agreement between raters in assigning a mastery level, needs to be also estimated. A simple method of estimating inter-rater reliability is to examine the association between the ratings of two examiners for the same group of children by estimating a correlation coefficient. This method, however, tends to overestimate inter-rater reliability as it merely evaluates *association* and not *agreement*. Specifically, it does not provide information about the agreement in the category to which children were assigned by the two independent examiners, thus spearman correlations provide a limited picture. Instead, estimating a Cohen's kappa coefficient provides a more accurate estimate of inter-rater reliability as it estimates agreement between raters beyond agreement due to chance.

Several forms of evidence are collected to evaluate the validity of the test inference. One form of evidence is content validity, which indicates the extent to which the content of the scale maps onto the skills or abilities that an educator/examiner hopes to assess. For instance, inclusion of a story on the Hindi reading test that is in accordance with stories in 2nd standard state textbooks for story content, length of sentence, length of story, and type of vocabulary supports the inference about children's ability to read a 2nd standard level text.

Commonly used methods to obtain empirical evidence for validity are *concurrent* validity, and *convergent-discriminant* validity[1]. Concurrent validity involves the examination of the magnitude of association between performance on the new tests (ASER tests) and performance on already established standardized tests that share a common inference and which serve as the criterion tests. Strong and positive correlations between the two tests indicate strong evidence of concurrent validity.

The second approach, convergent-discriminant validity indicates that there is a stronger correlation among related constructs (e.g. reading tests) than among less related constructs (e.g. math test and reading test). In addition, a test of basic reading ability is expected to correlate

[1] A third approach is to estimate *predictive* validity, i.e. the association of the test with skills of reading ability assessed at a future time point for the same group of children. This has not been explored due to the absence of longitudinal data.

more strongly with another test of basic reading ability than with a test of advanced reading ability. However, the correlation between tests of reading and math are also expected to be high as they draw on children's underlying cognitive ability (unless if the sample studied has a specific learning difficulty). Hence, the differences in magnitude of the correlation coefficients between tests of reading and tests of math are typically very small.

The reliability and validity of the ASER-reading (Hindi) and ASER-math testing tools is addressed through a series of studies. The first section reports the findings of Study 1 and Study 2 and is titled 'Reliability of the ASER Testing Tools'. Study 1 evaluates the reliability of decisions or decision consistency across repeated assessments and Study 2 evaluates inter-rater reliability. The second section reports the findings of Study 3 and is titled 'Validating the ASER Testing Tools.' Study 3 is based on data from the evaluation of a reading and math intervention program for which the ASER-reading (Hindi) and ASER-math tools have been used along with a battery of other literacy and math testing tools.

Reliability of the ASER Testing Tools

Study 1 and Study 2

Study Design and Participants

Study 1

Study 1 was conducted in Solan district, Himachal Pradesh. A total of 540 children from standards 1-5 were assessed by the same team of examiners on two different occasions with a lag of 2-3 days (see Table 1 for sample descriptives). In addition, for a sub-sample of 242 children, the order of test administration was changed such that 137 children were tested on reading then math on the first administration and were tested on math then reading on the second administration. One hundred and five children were tested on math then reading on the first administration and were tested on reading then math on the second administration. For 298 children the test order remained unchanged on both occasions and they were assessed on reading then math in keeping with the ASER survey methodology. Moreover, as per the ASER survey methodology all children were tested by a pair of examiners; in all there were 12 teams of examiners. The ASER 2008 test samples were used for Study 1.

Study 2

Study 2 was also conducted in Solan district, Himachal Pradesh. A total of 590 children from standards 1-5 were assessed by different pairs of examiners on two different occasions with a lag of 2-3 days (see Table 1). All assessments were conducted by a pair of examiners in keeping with the ASER survey methodology. There were 12 pairs of examiners and the pairs remained unchanged throughout the data collection process. Two pairs of examiners formed a team, thus forming 6 teams. Each team re-tested children tested by their team's pair. The ASER 2008 test samples were used for Study 2.

Analytic Plan

Test score distributions were examined for all tests for all studies.

Study 1

Differences in the first administration between the sub-groups of Study 1 were assessed using general linear models to ensure that there were no differences in sub-group selection.

To address the issue of order of test administration, we fit a set of general linear mixed regression models using SAS PROC MIXED for ASER-reading scores and for ASER-math scores separately. To assess the effect of order of test administration we created a series of dichotomous variables representing three groups: 1) children who were tested in the same order (reading then math) for test and retest, 2) children who were tested on reading then math on test and math then reading on retest, and 3) children who were tested on math then reading on test and vice versa on retest. By treating children who were tested in the same order as the reference group we assessed whether changing the order of test administration had any significant impact on math or reading performance.

To assess reliability we estimated a kappa[2] coefficient to index the decision consistency across repeated test administrations for the sample tested as part of Study 1. The kappa estimate varies from 0 to 1. A value of 0 indicates no improvement over chance and a value of 1 indicates maximal increase.

[2] The formula for Cohen's Kappa is $\kappa = \frac{p_o - p_c}{1 - p_c}$ where p_0 is the observed proportion of agreement between raters and p_c is the expected proportion of agreement, i.e. agreement due to chance. The coefficient kappa of 0 occurs when observed agreement can be exactly accounted for by chance and the coefficient kappa of 1 occurs when there is complete agreement between raters. Kappa can yield a negative value when there is less observed agreement than is expected by chance.

Table 1: Sample descriptives for the reliability studies

								Age
	n	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Female	M (SD)
		18%	20%	19%	21%	21%	52%	
Study 1	540	(98)	(110)	(104)	(113)	(115)	(304)	8.2 (1.8)
		19%	20%	21%	20%	20%	49%[a]	
Study 2	590	(115)	(120)	(122)	(117)	(116)	(267)	8.3 (1.9) [b]

[a]: For the Study 2 sample child gender is missing for 1 child

[b]: For the Study 2 sample child age is missing for 21 children

Study 2

To assess inter-rater reliability we also estimated a kappa coefficient to determine the degree of agreement between raters' classification of children into different mastery levels for reading and math. In estimating inter-rater reliability, two methods of estimating Kappa were employed: the simple Kappa evaluates for exact agreement, whereas the weighted Kappa assigns a higher weight to close misses (e.g. a rank of 2 vs. a rank of 3) than to misses that are further apart (e.g. a rank of 1 vs. a rank of 5) based on the assumption that a difference of adjacent ranks is less critical than a difference that is farther apart (Cohen, 1960; 1968). In other words, categorizing a 'story' level child to be at the 'para' level is less incorrect than categorizing a 'story' level child to be at the 'nothing' level.

Results

The distribution of scores for the ASER-reading test presented in Table 2 for the reliability study sample indicate that the distribution is somewhat skewed with a larger percentage of children categorized at the ‘story’ level[3]. The distribution of scores for the ASER-math test is more evenly distributed across the different math levels.

Table 3 indicates that there were no significant differences between the sub-groups of the test-retest reliability study sample on the first round of test administration (test) and the second round of test administration (retest) indicating that the sub-groups were randomly selected.

The taxonomy of general linear mixed regression models predicting for children’s ASER-reading and ASER-math test performance indicate that all children demonstrated an improvement in performance from the first testing session (test) to the second testing session (retest) (Table 4). However the lack of a significant effect of improvement by order of test administration indicates that the improvement was uniform across all sub-groups irrespective of order of administration. In other words order of test administration was not of consequence.

Since there was no effect of order of administration, the reliability of decisions was examined for the entire group of 540 children who were assessed by the same examiner on the two testing occasions. The test-retest correlation coefficients for the ASER-reading test for all children from standards 1-5 is .95 and for the ASER-math test is .90. More importantly the average kappa estimate for decision consistency across repeated test

[3] The reliability studies were conducted in Himachal Pradesh where basic levels of reading are known to be high based on previous ASER data and as reflected in the current findings.

administrations for the ASER-reading test is .76 and for the ASER-math test is .71. These estimates suggest ‘substantial’ level of agreement (Landis & Koch, 1977).

For Study 2, the inter-rater reliability estimated using Cohen’s Kappa for each team of two pairs of examiners ranged from .58 to .76 for the ASER-reading test and from .55 to .74 for the ASER-math test (Table 5). These simple Kappa estimates suggest ‘moderate’ agreement to ‘substantial’ agreement (Landis & Koch, 1977). The average and median Kappa estimate across all pairs of examiners is .64 and .63 respectively for the ASER-reading test and .65 and .66 respectively for the ASER-math test indicating ‘substantial’ agreement. The weighted Kappa estimate for the ASER-reading test ranges from .78-.84 and for the ASER-math test ranges from .71-.84. These weighted Kappa estimates range from ‘substantial’ agreement to ‘almost perfect’ agreement. The average and median weighted Kappa across all pairs of examiners is .82 and .81 respectively for the ASER-reading test and is .79 and .80 for the ASER-math test indicating ‘almost perfect’ agreement for the ASER-reading test and ‘substantial’ agreement for the ASER-math test (Landis & Koch, 1977).

The contingency table of scores and marginal distributions for the full sample are presented in Tables 6 and 7 for the ASER-reading test and the ASER-math test respectively for Study 1 and in Table 8 and 9 for the ASER-reading and the ASER-math test respectively for Study 2. The marginal distributions for the ASER-reading test are somewhat skewed, i.e. the ratings are concentrated at the higher levels of 4 and 5 and not similarly distributed across the full scale. Such skew tends to attenuate the Kappa estimates. Moreover, as noted in the general linear mixed regression models, children tend to perform better on retest than test even though their is high stability in relative ranking across the two testing sessions (*r*=.95 for ASER-reading and *r*=.90 for ASER-math). This upward shift in performance also impacts decision consistency and inter-rater reliability. As a result, the current estimates for simple and weighted Kappa for the reading tests are quite likely lower than would be estimated if the sample was more diverse in ability levels or were tested with a slightly longer lag.

Table 2: Score distributions for the ASER-reading and ASER-math test for Study 1 (decision-consistency) and Study 2 (inter-rater reliability)

		ASER-reading					ASER-Math				
		Nothing	Letter	Word	Passage	Story	Nothing	Single digit recognition	Double digit recognition	Subtracts	Divides
Study 1	All children: Test (n=540)	11% (61)	20% (106)	5% (26)	12% (66)	52% (281)	6% (31)	23% (126)	31% (165)	21% (111)	20% (107)
	All children: Retest (n=540)	10% (54)	19% (105)	5% (25)	10% (55)	56% (301)	5% (27)	21% (114)	29% (158)	20% (108)	25% (133)
	Order 1: Test (n=298)	9% (26)	22% (67)	4% (12)	12% (35)	53% (158)	5% (15)	23% (70)	30% (89)	22% (67)	19% (57)
	Order 1: Retest (n=298)	8% (25)	21% (63)	5% (14)	9% (27)	57% (169)	4% (11)	22% (66)	28% (84)	22% (65)	24% (72)
	Order 2: Test (n=137)	18% (24)	12% (16)	6% (8)	12% (17)	53% (72)	9% (12)	21% (29)	31% (43)	15% (20)	24% (33)
	Order 2: Retest (n=137)	15% (20)	15% (20)	4% (5)	12% (17)	55% (75)	9% (12)	18% (25)	28% (39)	16% (22)	28% (39)
	Order 3: Retest (n=105)	10% (11)	22% (23)	6% (6)	13% (14)	49% (51)	4% (4)	26% (27)	31% (33)	23% (24)	16% (17)
	Order 3: Retest (n=105)	9% (9)	21% (22)	6% (6)	10% (11)	54% (57)	4% (4)	22% (23)	33% (35)	20% (21)	21% (22)
Study 2	All children: test (n=590)	7% (42)	19% (115)	9% (51)	19% (110)	46% (272)	6% (33)	20% (120)	33% (196)	21% (126)	19% (115)
	All children: retest (n=590)	5% (32)	18% (109)	7% (44)	19% (112)	50% (293)	5% (28)	18% (109)	32% (187)	22% (131)	23% (135)
	Team1: Test (n=93)	3% (3)	20% (19)	5% (5)	18% (17)	53% (49)	5% (5)	24% (22)	33% (31)	24% (22)	14% (13)
	Team1: Retest (n=93)	3% (3)	20% (19)	1% (1)	17% (16)	58% (54)	3% (3)	22% (20)	33% (31)	15% (14)	27% (25)
	Team2: Test (n=94)	15% (14)	15% (14)	6% (6)	12% (11)	52% (49)	5% (5)	27% (25)	35% (33)	16% (15)	17% (16)
	Team2: Retest (n=94)	10% (9)	16% (15)	10% (9)	21% (20)	44% (41)	2% (2)	28% (26)	38% (36)	15% (14)	17% (16)
	Team3: Test (108)	4% (4)	19% (21)	7% (8)	29% (31)	41% (44)	4% (4)	14% (15)	29% (31)	25% (27)	29% (31)
	Team3: Retest (108)	3% (3)	18% (19)	6% (6)	21% (23)	53% (57)	3% (3)	16% (17)	26% (28)	25% (27)	31% (33)
	Team4: Test (88)	8% (7)	20% (18)	7% (6)	26% (23)	39% (34)	3 (3%)	24% (21)	34% (30)	23% (20)	16% (14)
	Team4: Retest (88)	5% (4)	23% (20)	3% (3)	20% (18)	49% (43)	5% (4)	15% (13)	30% (26)	35% (21)	16% (14)
	Team5: Test (101)	4% (4)	22% (22)	5% (50	7% (7)	62% (63)	5% (5)	21% (21)	33% (33)	23% (23)	18% (19)
	Team5: Retest (101)	4% (4)	20% (20)	4% (4)	12% (12)	60% (61)	6% (6)	18% (18)	31% (31)	27% (27)	19% (19)
	Team6_Test (106)	9% (10)	20% (21)	20% (21)	20% (21)	31% (33)	10% (11)	15% (16)	36% (38)	18% (19)	21% (22)
	Team6_Retest (106)	8% (9)	15% (16)	20% (21)	22% (23)	35% (370	9% (10)	14% (15)	33% (35)	17% (18)	26% (28)

Table 3: Sample mean, standard deviation and statistical test of differences between test and retest performance for the groups of children tested in different orders, i.e. Order 1 (reading then math on test and retest), Order 2 (reading then math on test and vice versa on retest), and Order 3 (math then reading on test and vice versa on retest).

	Order 1 (n=298)		Order 2 (n=137)		Order 3 (n=105)			
	M	SD	M	SD	M	SD	F-test	p-value
ASER-reading								
Test	3.78	1.49	3.71	1.60	3.68	1.51	0.22	0.80
Retest	3.85	1.48	3.78	1.57	3.81	1.48	0.09	0.91
ASER-math								
Test	3.27	1.16	3.24	1.27	3.22	1.12	0.09	0.92
Retest	3.41	1.18	3.37	1.31	3.32	1.15	0.19	0.83

Table 4: Taxonomy of fitted general linear mixed models examining for differences in children's ASER-reading and ASER-math test scores at test and retest by order of test administration controlling for standard, child age, and child gender (n=632)

	Model 1R	Model 2R	Model 3R	Model 4R	Model 1M	Model 2M	Model 3M	Model 4M
Fixed Effects: γ *(se)*								
Intercept	3.72***	3.60***	1.31***	1.31***	3.26***	3.07***	1.07***	1.07***
	(0.06)	(0.07)	(0.22)	(0.22)	(0.05)	(0.06)	(0.16)	(0.16)
Testing occasion		0.08***	0.08***	0.07**		0.13***	0.13***	0.13***
		(0.02)	(0.02)	(0.03)		(0.02)	(0.02)	(0.03)
Child age			0.03	0.03			0.02	0.02
			(0.03)	(0.03)			(0.02)	(0.02)
Child gender			-0.20**	-0.21**			-0.05	-0.06
			(0.08)	(0.08)			(0.05)	(0.05)
Standard			0.78***	077***			0.63***	0.63***
			(0.04)	(0.04)			(0.03)	(0.03)
Testing occasion x order of testing (Reading-Math then Math-Reading)				-0.03				-0.02
				(0.04)				(0.03)
Testing occasion x order of testing (Math-Reading then Reading-Math)				0.07				0.01
				(0.04)				(0.04)
Variance Components: σ^2 *(se)*								
Within-person variance	0.12***	0.12***	0.12***	0.12***	0.15***	0.14***	0.14***	0.14***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Initial status	2.14***	2.13***	0.84***	0.83***	1.23***	1.23***	0.38***	0.38***
	(0.12)	(0.12)	(0.05)	(0.05)	(0.07)	(0.07)	(0.03)	(0.03)
Goodness-of-fit								
-2LL	3071.0	3057.0	2495.9	2491.8	2832.5	2798.4	2143.1	2142.7
AIC	3077.0	3065.0	2509.9	2509.8	2838.5	2806.4	2157.1	2160.7

**p<.01, ***p<.001. Note: Gender is a dummy variable with 0 representing female

Table 5: Simple and weighted kappa estimates for inter-rater reliability

		Simple kappa	Weighted kappa
ASER-reading	Team 1 (n=93)	.58	.78
	Team 2 (n=94)	.61	.82
	Team 3 (n=108)	.65	.80
	Team 4 (n=88)	.68	.84
	Team 5 (n=101)	.76	.87
	Team 6 (n=106)	.58	.78
	All teams (n=590)	.64	.82
ASER-math	Team 1 (n=93)	.58	.78
	Team 2 (n=94)	.74	.83
	Team 3 (n=108)	.74	.84
	Team 4 (n=88)	.55	.71
	Team 5 (n=101)	.70	.81
	Team 6 (n=106)	.61	.77
	All teams (n=590)	.65	.79

Table 6: Score distributions and marginal distributions for the ASER-reading test for Study 1 (decision-consistency, n=540)

		Retest					
		1: Nothing level	2: Akshar level	3: Word level	4: Para level	5: Story level	Marginal distributions
Test	1: Nothing level	48[a] (9%)[b]	12 (2%)	0 (0%)	1 (0%)	0 (0%)	61 (11%)
	2: Akshar level	6 (1%)	90 (17%)	8 (2%)	2 (0%)	0 (0%)	106 (20%)
	3: Word level	0 (0%)	1 (0%)	11 (2%)	9 (2%)	5 (1%)	26 (5%)
	4: Para level	0 (0%)	1 (0%)	6 (1%)	35 (6%)	24 (4%)	66 (12%)
	5: Story level	0 (0%)	1 (0%)	0 (0%)	8 (2%)	272 (50%)	281 (52%)
	Marginal distributions	54 (10%)	105 (19%)	25 (5%)	55 (10%)	301 (56%)	540 (100%)

Note: Shaded cells indicate exact agreement in assigning ranks on test and retest
[a]: Observed cell value [b]: Cell value as a percentage of total observations

Table 7: Score distributions and marginal distributions for the ASER-math test for Study 1 (decision-consistency, n=540)

		Retest					
		1: Nothing level	2: Single digit recognition	3: Double digit recognition	4: Subtraction	5: Division	Marginal distributions
Test	1: Nothing level	21[a] (4%)[b]	10 (2%)	0 (0%)	0 (0%)	0 (0%)	31 (6%)
	2: Single digit recognition	6 (1%)	94 (17%)	25 (5%)	1 (0%)	0 (0%)	126 (23%)
	3: Double digit recognition	0 (0%)	10 (2%)	124 (23%)	21 (4%)	10 (2%)	165 (31%)
	4: Subtraction	0 (0%)	0 (0%)	7 (1%)	81 (15%)	23 (4%)	111 (21%)
	5: Division	0 (0%)	0 (0%)	2 (0%)	5 (1%)	100 (19%)	107 (20%)
	Marginal distributions	27 (5%)	114 (21%)	158 (29%)	108 (20%)	133 (25%)	540 (100%)

Note: Shaded cells indicate exact agreement in assigning ranks on test and retest
[a]: Observed cell value [b]: Cell value as a percentage of total observations

Table 8: Score distributions and marginal distributions for the ASER-reading test for Study 2 (inter-rater reliability, n=590)

		Retest					
		1: Nothing level	2: Akshar level	3: Word level	4: Para level	5: Story level	Marginal distributions
Test	1: Nothing level	25[a] (4%)[b]	17 (3%)	0 (0%)	0 (0%)	0 (0%)	42 (7%)
	2: Akshar level	6 (1%)	85 (14%)	19 (3%)	5 (1%)	0 (0%)	115 (19%)
	3: Word level	1 (0%)	5 (1%)	21 (4%)	22 (4%)	2 (0%)	51 (9%)
	4: Para level	0 (0%)	2 (0%)	4 (1%)	64 (11%)	40 (7%)	110 (19%)
	5: Story level	0 (0%)	0 (0%)	0 (0%)	21 (4%)	251 (43%)	272 (46%)
	Marginal distributions	32 (5%)	109 (18%)	44 (7%)	112 (19%)	293 (50%)	590 (100%)

Note: Shaded cells indicate exact agreement between raters on test and retest
[a]: Observed cell value [b]: Cell value as a percentage of total observations

Table 9: Score distributions and marginal distributions for the ASER-math test for Study 2 (inter-rater reliability, n=590)

		Retest					
		1: Nothing level	2: Single digit recognition	3: Double digit recognition	4: Subtraction	5: Division	Marginal distributions
Test	1: Nothing level	21[a] (4%)[b]	12 (2%)	0 (0%)	0 (0%)	0 (0%)	33 (6%)
	2: Single digit recognition	7 (1%)	84 (14%)	27 (5%)	0 (0%)	2 (0%)	120 (20%)
	3: Double digit recognition	0 (0%)	12 (2%)	146 (25%)	31 (5%)	7 (1%)	196 (33%)
	4: Subtraction	0 (0%)	1 (0%)	12 (2%)	86 (15%)	27 (5%)	126 (21%)
	5: Division	0 (0%)	0 (0%)	2 (0%)	14 (2%)	99 (17%)	115 (19%)
	Marginal distributions	28 (5%)	109 (18%)	187 (32%)	131 (22%)	135 (23%)	590 (100%)

Note: Shaded cells indicate exact agreement between raters on test and retest
[a]: Observed cell value [b]: Cell value as a percentage of total observations

Validating the ASER Testing Tools

Study 3

Study Design and Participants

As part of an evaluation of a reading and math intervention program the ASER-reading and ASER-math tests were administered along with a battery of tests of basic and advanced reading and math ability. The other tests included (a) the Fluency Battery, which is a test of basic reading ability similar to the ASER-reading test, (b) the Read India (RI) Literacy test, which is a paper-and-pencil test assessing basic and advanced reading and writing ability, and (c) the Read India (RI) Math test, which is also a paper-and-pencil test assessing basic and advanced math ability.

Several rounds of data were collected: (1) An initial pilot study (Pilot 1) with 256 children from standards 1-5, (2) a second pilot study (Pilot 2) conducted with 412 children from standards 1-5, (3) a baseline evaluation conducted in two districts in Bihar (n=8092) and Uttarakhand (n=7237) with children aged 5-16 from standards 1-8, and (4) a midline evaluation in Bihar conducted with 4807 children of ages 5-16 from standards 1-8 who were assessed on the ASER-reading and -math tests and the Fluency Battery. Data from all tests is available for Pilot 1, Pilot2, and the Bihar baseline study. For the Uttarakhand baseline[4] and the Bihar midline studies[5] data from the ASER-reading, ASER-math, and the Fluency Battery is alone available.

The READ INDIA Measures:

The Fluency Battery: The assessment of fluency is based on the premise that the ability to read fluently, i.e. with sufficient speed and accuracy is important to read well and to comprehend text. In fact, the fluent decoding of letters, letter combinations, words in list form, and words in connected text are important and robust correlates of early reading ability and comprehension. The automaticity of these lower-level skills ensures that limited cognitive resources such as attention and memory can be freed and allocated to the higher-level skills of meaning-making (LaBerge & Samuels, 1974; Perfetti, 1977, 1985). Hence, fluency measures, which are orally administered tests, are widely used to assess children's early reading ability in English and several other languages. The Fluency Battery was adapted from the Early Grade Reading Assessment (USAID, 2009) and the Dynamic Indicators of Basic Early Literacy Skills (University of Oregon Center on Teaching and Learning, 2002). It comprises 8 subtests:

[4] Data on the Read India Literacy and Math tests for the Uttarakhand sample are not used as several concerns remain about the credibility of these data.

[5] The Read India Literacy and Math tests were not administered for the Bihar midline evaluation.

1. Akshar Reading Fluency (ARF): indicates the speed and accuracy with which children read aloud randomly arranged akshars of the Hindi alphasyllabary in a span of one minute. The score is the number of akshars named correctly in 1 minute.
2. Barakhadi Reading Fluency (BRF): indicates the speed and accuracy with which children read aloud randomly arranged consonant-vowel (CV) akshar units in one minute. All units were represented by a single consonant /k/ so as not to confound this task with the Akshar Reading Fluency subtest. The score is the number of barakhadi units named correctly in 1 minute.
3. Word Reading Fluency (WRF): indicates the speed and accuracy with which children read aloud a list of one and two syllable words in one minute. The score is the number of words read correctly in 1 minute.
4. Nonword Reading Fluency (NWRF): indicates the speed and accuracy with which children read aloud a list of two syllable nonwords in one minute. The score is the number of nonwords read correctly in 1 minute.
5. Grade 1 Level Passage Reading Fluency (two passages, PRF): indicates the speed and accuracy with which children read aloud two standard 1 level passages comprising 4 sentences and 21 words. The score is an average of the two passages and indexes the number of words read correctly in 1 minute.
6. Grade 2 Level Passage Reading Fluency (two passages, SRF): indicates the speed and accuracy with which children read aloud two standard 2 level passages comprising 6 sentences and 59-63 words. The score is an average of the two passages and indexes the number of words read correctly in 1 minute.

7. Grade 1 level Passage Comprehension Questions (PCOMP): comprises two comprehension questions for each of the two standard 1 level passages. The score is the number of questions answered correctly.
8. Grade 2 level Passage Comprehension Questions (SCOMP): comprises four comprehension questions for each of the two standard 2 level passages. The score is the number of questions answered correctly.

The content of the Fluency Battery was drawn from prior ASER reading tests as the material has been extensively evaluated and piloted to ensure their grade and content appropriateness for the population of interest. There was no overlap in test content of the ASER reading tests and the Fluency Battery. Scores for the fluency reading subtests represent number of units (akshars/words/nonwords) read accurately in one minute and scores for the reading comprehension subtest represent number of questions correctly answered. Total administration time for the Fluency Battery is about 10 minutes. The median Cronbach's alpha estimates across the 5 samples ranged from .92 to .94 with a median Cronbach's alpha estimate of .93. Test-retest reliability coefficients for the subtests of the Fluency Battery ranged from .83 to .98. Since the association between the Fluency Battery sub-tests was high for all the 5 samples (*r*s range from .81 to .94) a single composite score of all the fluency sub-tests was created by taking an average.

The Literacy and Math Written Tests: A more traditional format of written tests of reading and math were developed to assess higher level reading, writing, and math skills. These tests were drawn from extensively piloted Urdu reading and math tests for use in Pakistan (Andrabi, Das, Khwaja, Farooqi, & Zajonc, 2002) and from the math tests of the TIMMS (http://nces.ed.gov/timss/). A few math items were also drawn from the ASER. The format of the questions on the Pakistan Urdu test were used as a reference to develop test items for Hindi, e.g. format of the reading vocabulary items, cloze sentences, maze passage, etc. Care was taken to ensure that the content and item formats were appropriate for use in Hindi and aligned to the Bihar and Uttarakhand language and math curriculum, the two states that are part of the Read India evaluation study.

In order to ensure that the test items were appropriate for grade level, separate tests with overlapping content were designed for grades 1-2 and grades 3-5. The RI Literacy test for grades 1-2 requires about 15 minutes to administer and for grades 3-5 requires about 20 minutes to administer. The RI Math test for grades 1-2 requires about 10 minutes to administer and for grades 3-5 requires about 20 minutes to administer. Reliability based on internal consistency was estimated in two ways (1) treating each item on the test as an individual item, and (2) treating each questions category on the test as an individual item thereby reducing the count of the number of items on the test. For the tests for standards 1-2, the median Cronbach's alpha estimate for the first approach was .93 for the RI Literacy test and .93 for the RI Math test. The median Cronbach's alpha estimate for the second approach was .86 for the RI Literacy test and .86 for the RI Math test. For the tests for standards 3-5, the median Cronbach's alpha estimate for the first approach was .93 for the RI Literacy test and .94 for the RI Math test. The median Cronbach's alpha estimate for the second approach was .88 for the RI Literacy test and .90 for the RI Math test (see Abdul Latif Jameel Poverty Action Lab, Pratham, & ASER, 2009 for a detailed evaluation of all READ INDIA tests).

Analytic Plan

Study 3

To assess concurrent validity, we estimated the degree of association between the ASER-reading test and the Fluency Battery and the RI Literacy test using Spearman correlation coefficients. We expected the ASER-reading tests to be strongly correlated with both tests but we expected correlations of higher magnitude between the ASER-reading test and the Fluency Battery than with the RI Literacy test as the former two tests share a common inference about children's basic reading ability and are in the oral format. The association of the ASER-reading test and the RI Literacy test also helps us understand the relationship between literacy tests in the oral and written format. For the ASER-math test we estimated the degree of association between the ASER-math test and the RI Math test using Spearman correlation coefficients.

To assess convergent-discriminant validity, we evaluated the differences in the estimated degree of association of the ASER-reading test with the other tests of literacy versus with the tests of math. Similarly, we evaluated the differences in the estimated degree of association of the ASER-math test with the test of math versus the tests of literacy. These were estimated separately for the 5 samples – Pilot 1, Pilot 2, the Bihar baseline, the Uttarakhand baseline, and the Bihar midline sample.

Comparing Performance on the ASER and the Fluency Battery: A Closer Look

Since the ASER-reading test and the Fluency Battery are tests of early reading ability, two additional sets of analysis were conducted for these two tests to better understand the appropriateness of the cut-off criteria used for the ASER-reading test. First, the fluency rates were examined for children classified at different reading levels based on the ASER-reading test. Second, the percentage of children on the Fluency Battery who read less than 3 akshars/words and more than 3 akshars/words was calculated. These percentages were calculated for each reading level of the ASER-reading test thus permitting an evaluation of decisions based on a short test such as the ASER that has only 5 items (akshars) on the akshar reading subtest and 5 items (words) on the word reading subtest versus the 52 akshars on the Akshar Reading Fluency subtest and 52 words on the Word Reading Fluency subtest, albeit with a time limit of 1 minute. These sets of analysis allow evaluating agreement between decisions across tests that are administered independently yet designed to assess the same abilities or skills.

Results

The concurrent validity coefficients presented in Tables 10a-10c indicate that the ASER-reading test is very highly correlated with the Fluency Battery. The magnitude of the correlation coefficients range from .90 to .94. In addition, these coefficients indicate that the ASER-reading test is more strongly correlated with the Fluency Battery than it is with the ASER-math test. This pattern is evident across all the validity studies, i.e. pilot 1, pilot 2, the Bihar baseline, the Uttarakhand baseline, and the Bihar midline. In addition, the correlation coefficients in Tables 11a-11d indicate that the ASER-reading test is more strongly correlated with the RI Literacy test than with the math tests except for standard 1-2 for the Bihar baseline study. In the latter case, the ASER-reading test is more strongly associated with the ASER-math test than with the RI Literacy test.

Table 10a: Validity coefficients: Pilot 1, standards 1-5 (n=256) is below the diagonal and Pilot 2, standards 1-5 (n=390) is above the diagonal

	Fluency Battery	ASER-reading	ASER-math
Fluency Battery	--	.90	.80
ASER-reading	.91	--	.77
ASER-math	.81	.76	--

Note: All correlation coefficients are significant at p<.001

Table 10b: Validity coefficients for the Bihar baseline, standards 1-8 (n=7747) is below the diagonal and for the Uttarakhand baseline, standards 1-8 (n=7160) is above the diagonal

	Fluency Battery	ASER-reading	ASER-math
Fluency Battery	--	.94	.84
ASER-reading	.91	--	.82
ASER-math	.87	.87	--

Note: All correlation coefficients are significant at p<.001

Table 10c: Validity coefficients for the Bihar midline, standards 1-8 (n=4725)

	Fluency Battery	ASER-reading	ASER-math
Fluency Battery	--		
ASER-reading	.94	--	
ASER-math	.87	.87	--

Note: All correlation coefficients are significant at p<.001

The concurrent validity coefficients presented in Tables 11a-11d indicate that the ASER-math test, which is a test of basic math ability is moderately strongly to strongly correlated with the RI Math test, which is a test of basic to advanced math ability. The magnitude of the correlation coefficients range from .74 to .90. Moreover, for standards 1-2 in the two pilot studies, the ASER-math test is equally strongly correlated with the literacy tests as it is with the RI math test. However, the ASER-math test is more strongly correlated with the RI Math test than with the literacy tests for all standards for the Bihar baseline and for standards 3-5 in the two pilot studies.

Table 11a: Validity coefficients for Pilot 1 for grades 1-2 (n=96)

	FBT	ASER-reading	RI Literacy	ASER-math
FBT	--			
ASER-reading	.92	--		
RI Literacy	.88	.87	--	
ASER-math	.79	.78	.76	--
RI Math	.78	.79	.86	.79

Note: All correlation coefficients are significant at p<.001

Table 11b: Validity coefficients for Pilot 1 for grades 3-5 for the RI Literacy test (n=94) is below the diagonal and for the RI Math (n=72) is above the diagonal

	FBT	ASER-reading	RI Literacy/Math	ASER-math
FBT	--	.88	.74	.79
ASER-reading	.86	--	.69	.74
RI Literacy/Math	.89	.81	--	.90
ASER-math	.77	.73	.76	--

Note: (a) All correlation coefficients are significant at p<.001 (b) The RI Literacy and Math tests were administered to different groups of children, hence the validity coefficients are estimated separately for the Literacy and Math tests.

Table 11c: Validity coefficients for Bihar baseline: Grades 1-2 (n=3818) is below the diagonal and grades 3-8 is above the diagonal (n=3035)

	FBT	ASER-reading	RI Literacy	ASER-math	RI Math
FBT	--	.82	.81	.73	.76
ASER-reading	.76	--	.76	.72	.72
RI Literacy	.68	.65	--	.73	.83
ASER-math	.70	.72	.64	--	.79
RI Math	.69	.66	.75	.74	--

Note: All correlation coefficients are significant at p<.001

Table 11d: Pilot 2: Grades 1-2 (n=171) is below the diagonal and grades 3-5 (n=220) is above the diagonal

	FBT	ASER-reading	RI Literacy	ASER-math	RI Math
FBT	--	.85	.83	.69	.70
ASER-reading	.82	--	.77	.65	.67
RI Literacy	.88	.82	--	.72	.82
ASER-math	.71	.68	.74	--	.80
RI Math	.75	.72	.81	.74	--

Note: All correlation coefficients are significant at p<.001

How did children at the different ASER-reading levels perform on the Fluency Battery?

The descriptive statistics for fluency rates for children at the different ASER-reading levels presented in Table 12 indicates that reading fluency rates increase with the increasing ASER-reading levels. In other words, children categorized at the standard 2 story reading level (level 5) read the akshars, barakhadi, words, nonwords, and words in connected text with greater speed and accuracy then children classified at any of the lower levels of reading on the ASER-reading test. For instance, the fluency rates for akshars averaged across the four samples are about 2 for children at the ‘nothing’ level, about 17 for children at the ‘akshar’ level, 32 for children at the ‘word’ level, 45 for children at the ‘para’ level, and 62 for children at the ‘story’ level. These increasing fluency rates with higher ASER-reading levels are reflected in the strong validity coefficients between the ASER-reading test and the Fluency Battery noted earlier.

Given that the ASER-reading levels are mutually exclusive categories, children classified at the ‘akshar’ level are seen to demonstrate competency at the akshar level but not at the word level, and so on. It follows then that children at the ‘nothing’ level should perform poorly on the akshar reading fluency subtest and children at the 'akshar' level should perform poorly on the word reading fluency subtest and so on. Average performances presented in Table 12 substantiate this claim. For instance, averaging across the 4 samples, children classified at the 'nothing' level demonstrate akshar fluency rates of 2 akshars, children classified at the 'akshar' level demonstrate word fluency rates of 3 words, children classified at the 'word' level demonstrate standard 1 level oral fluency rates of 25 words, children classified at the standard 1 passage level demonstrate standard 2 level oral fluency rates of 44 words[6].

The ASER akshar and word reading subtests are extremely short tests that comprise only 5 items. As a result it is possible that children can be misclassified due to item sampling error. To evaluate the efficacy of such a short test the percentage of children who identified no akshars/words, who identified less than 4 akshars/words and who identified >4 akshars/words on the Akshar and Word Reading Fluency subtests was calculated. This enabled comparing children's performance on the ASER akshar and word reading subtests with performance on the akshar and word reading fluency subtests that comprise all the akshars of the Hindi alphasyllabary and a substantially larger number of words.

Results presented in Table 13 indicate that of the children classified at the 'nothing' level 82% of the children in Uttarakhand, 94% of the children in the Bihar baseline study, and 95% of the children in the Bihar midline study could not correctly identify 4 or more akshars on the Akshar reading fluency subtest. Of the children classified at the 'akshar' level 96% of the children in

[6] Much variation is noted for the fluency rates as some children demonstrate high fluency rates despite being categorized at lower levels of reading. A few instances of misclassification referred to as decision inconsistency is to be expected. However, the percentage of these misclassifications is on the low side (see next set of analysis). This warrants further examination if the ASER-reading tests are to be used for diagnostic purposes or for making decisions at the individual rather than the group level.

Uttarakhand, 80% of the children in the Bihar baseline study, and 85% of the children in the Bihar midline study could in fact correctly identify 4 or more akshars.

Of the children classified at the ‘word’ level 98% of the children in Uttarakhand, 87% of the children in the Bihar baseline study, and 96% of the children in the Bihar midline study did correctly read 4 or more words correctly. This is a high level of consistency across the two tests. However there are children who were classified at the ‘nothing’ level who correctly read more than 3 akshars in one minute on the Akshar Reading Fluency subtest and there were children classified at the ‘akshar’ level who correctly read more than 3 words (Table 13). Further examination of the fluency rates for these decision inconsistencies indicates that although the children categorized at the ‘nothing’ level read 4 or more akshars correctly on the Akshar Reading Fluency subtest, they demonstrated low rates of fluency in comparison to their counterparts who were categorized at the ‘akshar’ level (Table 14 presents descriptives and Figures 1a-1c present score distributions). Similarly, children categorized at the ‘akshar’ level read 4 or more words correctly on the Word Reading Fluency subtest, they demonstrated low rates of fluency in comparison to their counterparts who were categorized at the ‘word’ level (Table 15 presents descriptives and Figures 2a-2c present score distributions).

Table 12: Descriptive statistics for the Fluency Battery for children classified at different ASER-reading levels

	Fluency Battery								
ASER-Reading Level	Akshar Reading Fluency	Barakhadi Reading Fluency	Word reading fluency	Nonword reading fluency	Grade 1 level passage reading fluency	Grade 2 level passage reading fluency	Oral reading fluency	Grade 1 level comprehension	Grade 2 level comprehension
Uttarakhand Baseline									
Nothing Level (n=1775)	1.90 (4.2)	0.59 (2.1)	0.09 (0.60)	0.03 (0.52)	0.17 (1.43)	0.12 (1.17)	0.15 (1.18)	0.06 (0.32)	0.03 (0.27)
Akshar Level (n=1726)	22.65 (12.34)	8.46 (10.64)	4.41 (5.70)	1.43 (2.85)	7.23 (11.89)	6.67 (9.86)	6.95 (10.68)	0.97 (1.21)	0.87 (1.76)
Word Level (n=470)	37.40 (13.19)	28.50 (17.19)	17.90 (11.44)	8.18 (6.73)	29.98 (22.27)	26.37 (17.66)	28.17 (19.58)	2.50 (1.19)	3.91 (2.11)
Para Level (n=847)	45.05 (14.75)	39.47 (18.27)	28.14 (14.37)	13.28 (8.75)	52.49 (27.71)	44.60 (22.51)	48.54 (24.36)	3.12 (1.02)	5.27 (2.06)
Story Level (n=2361)	64.63 (21.33)	68.28 (22.28)	67.22 (25.68)	38.28 (18.36)	116.20 (42.39)	101.69 (36.36)	108.95 (38.15)	3.61 (0.65)	6.59 (1.44)
Bihar Baseline									
Nothing Level (n=4078)	1.10 (4.88)	0.84 (5.37)	0.39 (4.26)	0.28 (3.71)	0.38 (6.56)	0.38 (5.51)	0.38 (5.92)	0.02 (0.26)	0.03 (0.38)
Akshar Level (n=1564)	15.41 (13.54)	9.33 (13.25)	3.88 (8.46)	2.22 (7.36)	4.11 (11.05)	4.41 (11.17)	4.26 (10.53)	0.29 (0.81)	0.32 (1.08)
Word Level (n=592)	30.86 (14.87)	26.85 (19.42)	18.23 (14.38)	11.30 (11.33)	22.30 (21.16)	20.63 (19.17)	21.46 (18.92)	1.50 (1.43)	1.48 (1.98)
Para Level (n=836)	45.50 (20.15)	47.20 (23.69)	37.63 (21.43)	23.44 (16.04)	53.78 (34.19)	46.19 (28.99)	49.99 (29.00)	2.76 (1.28)	3.82 (2.42)
Story Level (n=1796)	63.37 (24.48)	71.16 (28.05)	62.82 (27.99)	40.93 (20.93)	93.04 (48.00)	78.87 (37.53)	85.67 (39.30)	3.49 (0.86)	5.98 (2.05)
Bihar Midline									
Nothing Level (n=2193)	1.03 (2.32)	0.73 (2.53)	0.07 (1.67)	0.03 (0.88)	0.11 (3.08)	0.09 (2.89)	0.10 (2.96)	0.02 (0.18)	0.02 (0.23)
Akshar Level (n=1135)	12.64 (10.67)	5.41 (6.25)	2.10 (3.27)	0.70 (1.62)	2.43 (5.35)	2.16 (4.63)	2.29 (4.87)	0.52 (0.93)	0.25 (0.89)
Word Level (n=337)	30.61 (11.54)	21.42 (14.07)	12.93 (8.55)	7.07 (5.31)	18.88 (15.46)	16.72 (11.98)	17.80 (13.37)	1.78 (1.35)	2.50 (2.11)
Para Level (n=595)	42.70 (13.17)	40.74 (18.20)	28.27 (13.69)	15.37 (8.95)	46.33 (25.27)	39.66 (23.89)	42.99 (23.04)	2.82 (1.09)	4.05 (2.30)
Story Level (n=1352)	65.58 (21.60)	73.34 (24.22)	62.89 (25.35)	35.51 (17.00)	104.10 (41.39)	90.13 (36.67)	97.11 (37.96)	3.51 (0.76)	6.15 (1.91)
Pilot 1									
Nothing Level (n=34)	2.21 (2.10)	2.18 (4.07)	0.32 (1.09)	0.06 (0.34)	0.03 (0.17)	0 (0)	0.01 (0.09)	0.03 (0.17)	0 (0)
Akshar Level (n=66)	15.63 (10.48)	9.35 (11.69)	3.56 (5.25)	1.35 (2.59)	5.16 (12.26)	4.95 (9.73)	5.05 (10.47)	0.50 (0.98)	0.45 (1.21)
Word Level (n=23)	29.68 (11.55)	28.65 (16.03)	17.18 (12.40)	7.17 (5.44)	30.73 (28.75)	25.06 (19.03)	27.90 (23.12)	2.30 (1.11)	2.70 (2.38)
Para Level (n=40)	45.79 (14.51)	47.34 (19.57)	32.11 (15.49)	16.67 (8.11)	54.70 (27.70)	45.15 (21.34)	49.52 (22.95)	3.13 (0.69)	4.65 (1.76)
Story Level (n=93)	55.39 (17.48)	64.76 (18.21)	56.83 (19.01)	30.32 (11.79)	94.13 (34.10)	79.33 (27.94)	87.22 (28.91)	3.41 (0.77)	5.85 (2.06)

Table 13: Percentage of decision consistencies and inconsistencies across the Fluency Battery and the ASER-reading test.

	Uttarakhand Baseline					Bihar Baseline					Bihar Midline				
	Nothing Level (n=1775)	Akshar Level (n=1726)	Word Level (n=470)	Para Level (n=847)	Story Level (n=2361)	Nothing Level (n=4078)	Akshar Level (n=1564)	Word Level (n=592)	Para Level (n=836)	Story Level (n=1796)	Nothing Level (n=4078)	Akshar Level (n=1564)	Word Level (n=592)	Para Level (n=836)	Story Level (n=1796)
Akshar reading fluency															
0-3 akshars	81.98%	3.65%	0.00%	0.00%	0.00%	93.55%	20.46%	3.38%	2.03%	0.50%	94.80%	15.42%	0.59%	0.17%	0.00%
> 3 akshars	18.03%	96.35%	100.00%	100.00%	100.00%	6.45%	79.54%	96.62%	97.97%	99.50%	5.20%	84.58%	99.41%	99.83%	100.00%
Word reading fluency															
0-3 words	99.38%	57.99%	2.34%	0.71%	0.04%	98.62%	72.18%	13.17%	2.87%	0.72%	99.59%	80.17%	4.45%	0.17%	0.07%
>3 words	0.62%	42.00%	97.66%	99.29%	99.96%	1.37%	27.81%	86.82%	97.13%	99.28%	0.41%	19.82%	95.55%	99.83%	99.93%

Table 14: Descriptive statistics of akshar fluency rates for children whose akshar fluency rates are 4 or more and were categorized at the 'nothing' level or ‘akshar’ level on the ASER-reading test

		M	SD
Uttarakhand Baseline	Nothing level (n=320)	8.76	6.21
	Akshar level (n=1663)	23.46	11.84
Bihar Baseline	Nothing level (n=263)	13.75	13.91
	Akshar level (n=1244)	19.19	12.66
Bihar Midline	Nothing level (n=114)	7.69	7.02
	Akshar level (n=960)	14.67	10.39

Table 15: Descriptive statistics of akshar and word fluency rates for children whose word fluency rates are 4 or more and were categorized at the ‘akshar’ level or ‘’word’ level on the ASER-reading test

		Akshar fluency rates		Word fluency rates	
		M	SD	M	SD
Uttarakhand Baseline	Akshar level (n=725)	31.24	10.79	9.06	6.17
	Word level (n=459)	37.66	13.15	18.28	11.32
Bihar Baseline	Akshar level (n=435)	26.73	13.46	12.94	11.9
	Word level (n=514)	33.28	13.55	20.9	13.56
Bihar Midline	Akshar level (n=225)	23.93	9.2	7.2	4.1
	Word level (n=332)	31.24	11.22	13.45	8.39

Figure 1: Distribution of akshar reading fluency rates for children whose akshar fluency rates are 4 or more and were categorized at the 'nothing' level or 'akshar' level on the ASER-reading test

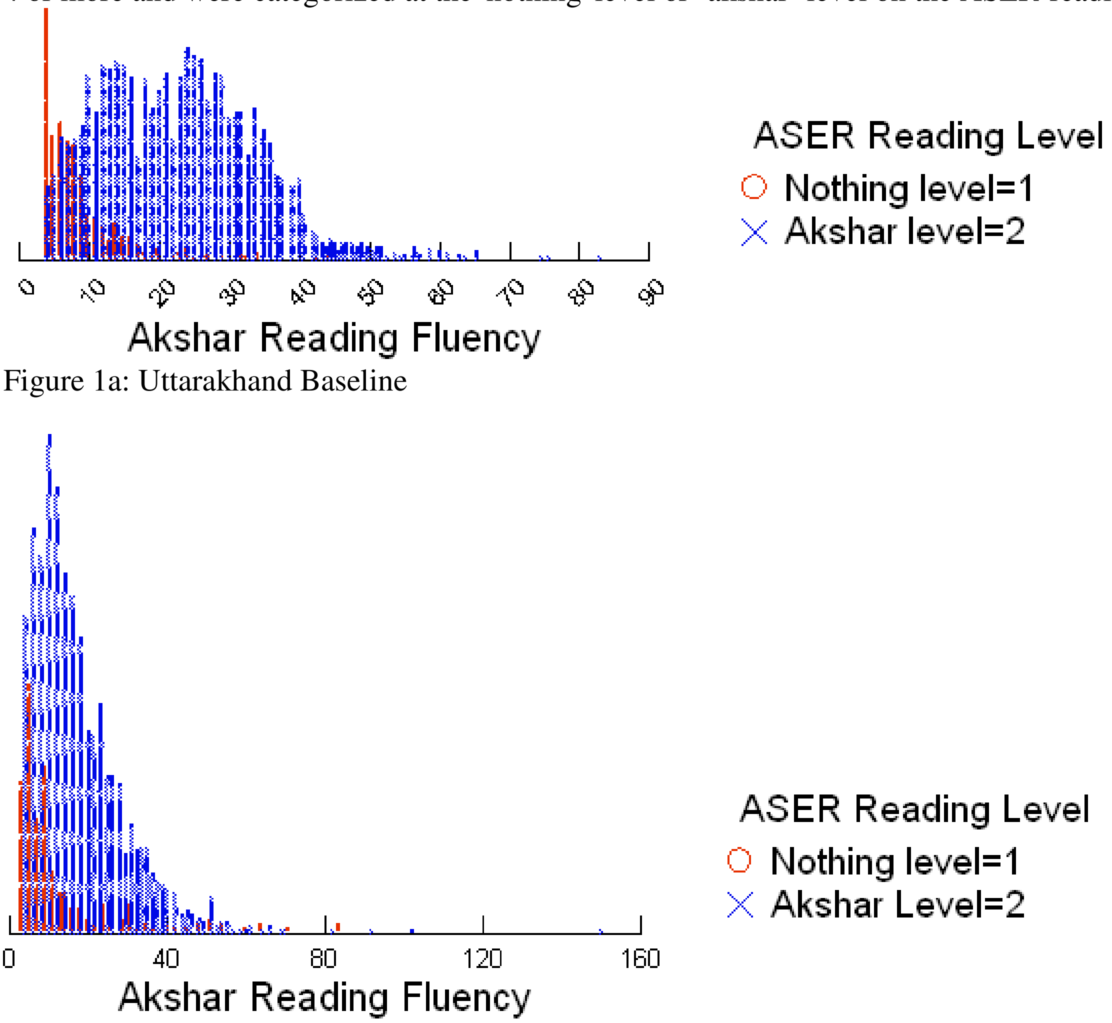

Figure 1a: Uttarakhand Baseline

Figure 1b: Bihar Baseline

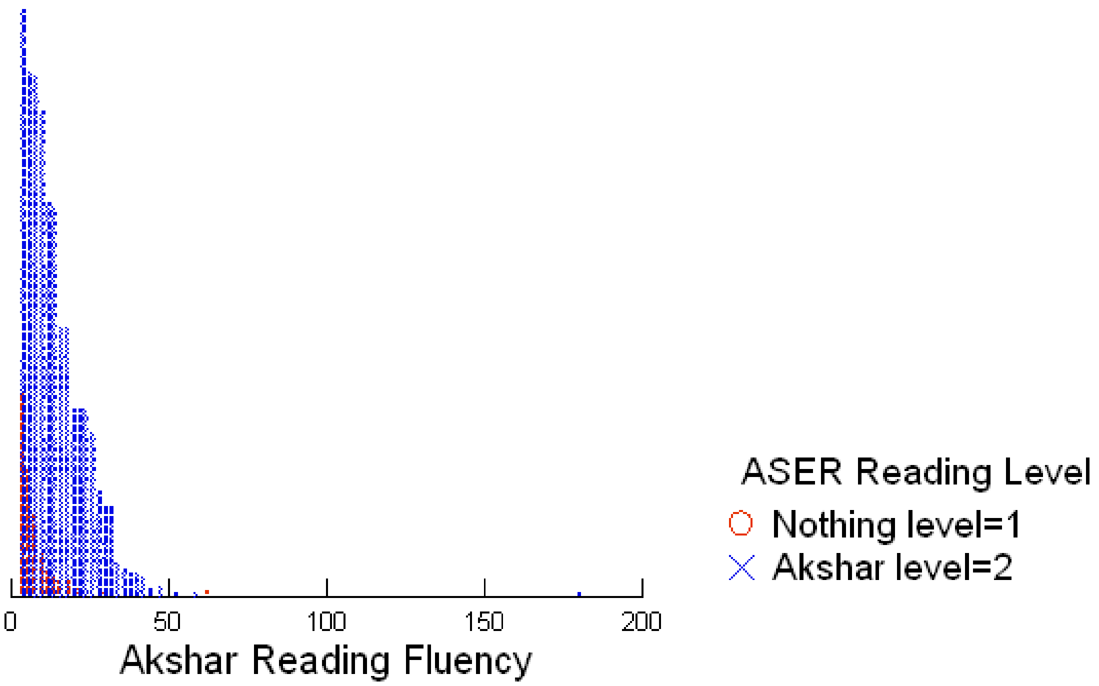

Figure 1c: Bihar Midline

Figure 2: Distribution of scores of akshar and word fluency rates for children whose word fluency rates are 4 or more and were categorized at the ‘akshar’ level or ‘’word’ level on the ASER-reading test

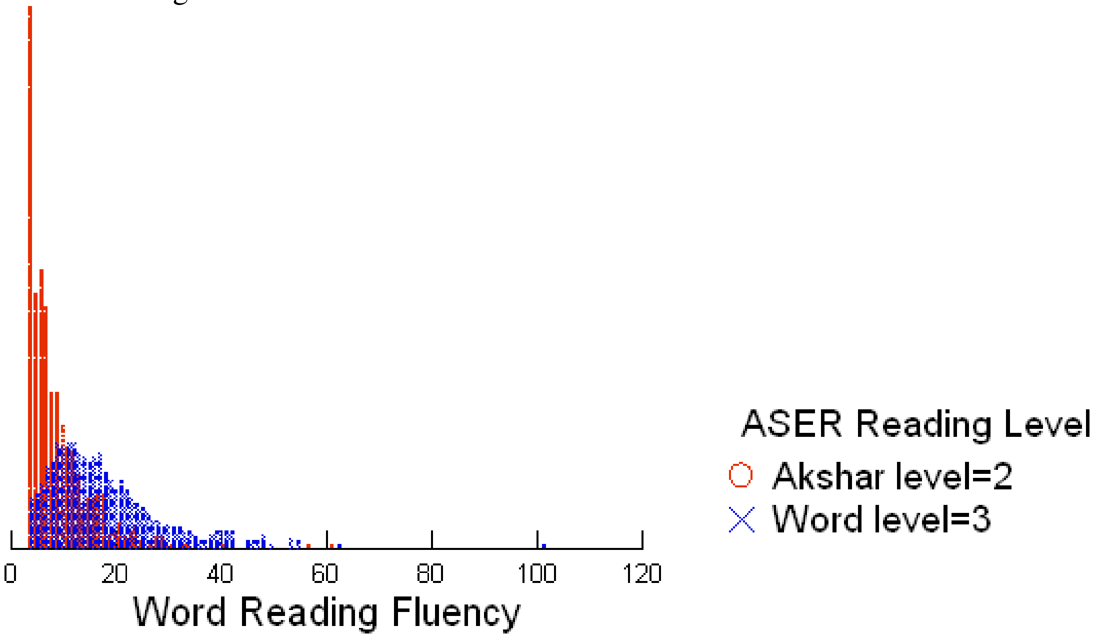

Figure 2a: Uttarakhand Baseline

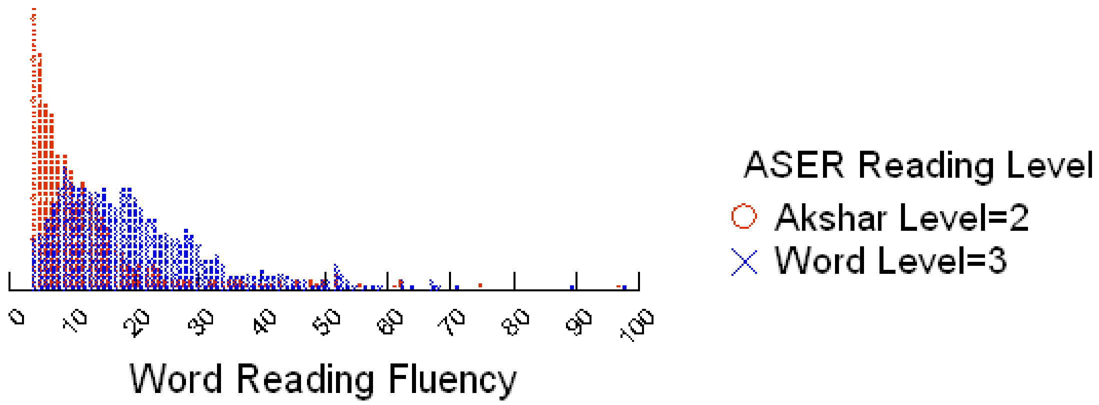

Figure 2b: Bihar Baseline

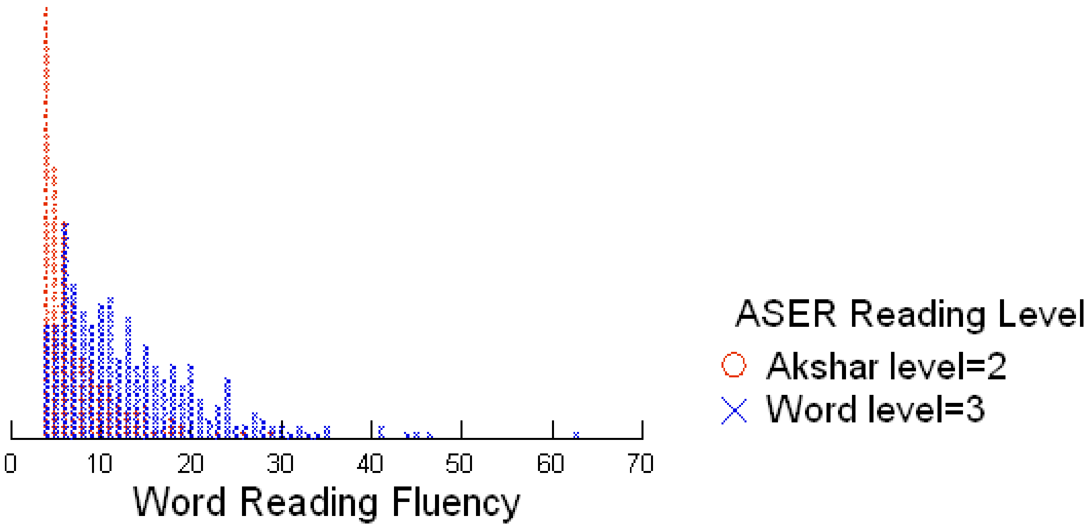

Figure 2c: Bihar Midline

Reliability and Validity of the ASER Testing Tools

Discussion

The ASER-reading and ASER-math tests are simple, quick, easy to administer and used primarily to obtain school-level and district-level data about children’s foundational reading skills and basic math ability. The findings based on a series of studies reported in the present paper provide favorable empirical evidence for the reliability and validity of these tests.

Specifically, the findings indicate substantial reliability of decisions across repeated measurements, satisfactory inter-rater reliability and favorable evidence for concurrent and convergent-discriminant validity.

Compelling evidence for the validity of the ASER tests is illustrated in (a) the very strong associations of the ASER-reading test with the concurrently administered Fluency Battery, which like the ASER-reading test assesses foundational reading skills, (b) the stronger association of the ASER-reading test with the Fluency Battery than with the RI Literacy test, which unlike the Fluency Battery also assesses advanced reading and writing ability, (c) the stronger association of the ASER-reading test with the Fluency Battery and the RI Literacy test than with the math tests, and (d) the stronger association of the ASER-math test with the RI Math test than with the tests of literacy.

Additional comparisons of the decision consistency between the ASER-reading test and the Fluency Battery indicate that there is a high level of consistency across the two tests at the ‘nothing’, ‘akshar’, and ‘word’ level. Although there were some inconsistencies with children at the ‘nothing’ level correctly reading 4 or more ‘akshars’ on the Akshar Reading Fluency subtest and with children at the ‘akshar’ level correctly reading 4 or more words on the Word Reading Fluency subtest, the respective fluency rates were clustered at the lower end of the continuum. Moreover, given that the ASER reading levels are mutually exclusive categories it implies that children who demonstrate competency at the akshar level do not demonstrate competency at the word or any other higher level. As a result, the fluency rates of children at the akshar level are bound to be lower than the fluency rates of children who are classified at the word or higher level. This expectation is supported by the data and is in keeping with the viewpoint that fluency in reading words in connected text requires fluency at the levels of smaller units such as letters (akshars) and letter combinations (barakhadi) (Foulin, 2005, Wolf & Katzir-Cohen, 2001). Consequently, an important instructional implication of this finding is that children categorized at the 'akshar' level are demonstrating 'minimal' mastery as opposed to 'complete' mastery of akshar knowledge and need to further improve their akshar knowledge if they are to successfully decode words in list form or connected text. Similarly, children classified at the 'word' level are demonstrating 'minimal' mastery of their decoding knowledge and need to further improve their decoding skills in order to fluently read and comprehend words in connected text.

Overall, the series of studies reported in the present paper provide favorable evidence for the reliability and validity of the ASER-reading and the ASER-math tests. Additional work with more diverse samples including an evaluation of alternate form reliability will help provide a more comprehensive picture of the psychometric properties of the ASER tests.

Finally, although the association between the ASER-reading test and the Fluency Battery is very strong and they both assess foundational reading skills, the decision to use any one of the tests should be based on considerations of the purpose of testing and the nature of information desired. The ASER-reading test provides information about children's reading levels in mutually exclusive ordinal ranks, whereas the Fluency Battery provides information about children's reading in terms of fluency at different levels of reading (akshars/words read correctly in one minute). Hence, both tests have their merits depending on the purpose of assessment. For instance, using the Fluency Battery along with the ASER-reading test in the evaluation of the

READ INDIA intervention program enables the assessment of children's progress in reading within and across reading levels. On the other hand, using the ASER-reading test for the nationwide ASER survey provides a reliable and valid snapshot of children's foundational reading skills in a simple, quick, cost-effective manner with results that are easy for policy makers, educators, and parents to understand and which are available in the same academic year.

References

Abdul Latif Jameel Poverty Action Lab [J-PAL], Pratham, & ASER, (2009). *Evaluating READ INDIA: the development of tools for assessing Hindi reading and writing ability and math skills of rural Indian children in grades 1-5.* Unpublished manuscript: J-PAL, Chennai, India.

Andrabi, Das, Khwaja, Farooqi, & Zajonc. (2002). Test feasibility survey Pakistan: Education Sector.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 71*, 213-220.

Cronbach, L.J. (1971). Test validation. In R.L.Thorndile (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Foulin, J.N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Reading and Writing, 18*, 129-155.

LaBerge, F., & Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293-323.

Landis, J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Perfetti, C.A. (1977). Literacy comprehension and fast decoding. Some psycholinguistic prerequisites for skilled reading comprehension. In J.T.Guthrie (Ed.), *Cognition, curriculum, and comprehension* (pp. 20-41). Neward, DE: International Reading Association.

Perfetti, C.A. (1985). *Reading ability.* London: Oxford.

Pratham (2005). Annual Status of Education Report (ASER). Retrieved July 1, 2009 from the World Wide Web: http://asercentre.org/asersurvey/aser05.php

Pratham (2006). Annual Status of Education Report (ASER). Retrieved July 1, 2009 from the World Wide Web: http://asercentre.org/asersurvey/aser06.php

Pratham (2007). Annual Status of Education Report (ASER). Retrieved July 1, 2009 from the World Wide Web: http://asercentre.org/asersurvey/aser07.php

Pratham (2008). Annual Status of Education Report (ASER). Retrieved July 1, 2009 from the World Wide Web: http://asercentre.org/asersurvey/aser08/pdfdata/aser08.pdf

Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of criterion-referenced tests: a decision-theoretic formulation. *Journal of Educational Measurement, 11*(4), 263-267.

Traub, R.E., & Rowley, G.L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement, 4*(4), 517-545.

Trends in International Mathematics and Science Study [TIMSS] (2003). *Mathematics Items: Grade 4.* Retrieved February 1, 2008 from the World Wide Web: http://nces.ed.gov/TIMSS/pdf/TIMSS4_Math_Items.pdf

USAID (2009). Early grade reading assessment (EGRA). Retrieved July 1, 2009 from the World Wide Web: http://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=95

University of Oregon Center on Teaching and Learning. (2002). *Dynamic indicators of basic early literacy skills [DIBELS]: Analysis of reading assessment measures*. Retrieved February 1, 2008 from the World Wide Web: http://dibels.uoregon.edu/techreports/dibels_5th_ed.pdf

Wagner, D.A. (2003). Smaller, quicker, cheaper: alternative strategies for literacy assessment in the UN Literacy Decade. *International Journal of Educational Research, 39*, 293-309.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5*(3), 211-239.