ASER
Evidence for Action

Bihar Elementary School Study

May 2014

Provisional

unicef
unite for children

असर ASER
Evidence for Action

Bihar Elementary School Study
May 2014

All photos taken by Pratham/ASER Centre team as they visited schools and villages.

Preface

Conducted in May 2014, the "Bihar Elementary School Study" (BESS) was a joint effort of ASER Centre/Pratham, SCERT (Bihar Government) and UNICEF. The study was designed to collect information on learning outcomes of students studying in primary schools (Std 2, 4 and 6) in all 38 districts of the state. This study focussed on basic and fundamental learning as well as on grade level learning achievements. Also the entire process had a major capacity building component built-in so that the evidence generated could potentially help the education system to plan how students' learning could improve.

The "BESS May 2014 Technical Report" is an additional report to the BESS May 2014 report. The "Technical Report" describes the methodologies along with other features, which enable a robust and quality data collection process to be put in place. The purpose of writing this report is to find out whether the general findings based on students results can be validated along with the findings of the psychometric report which checks the student performance in accordance with the tool reliability validity and difficulty. To ensure whether that the tools utilized for the assessment are reliable and valid, a technical psychometric report has been constructed. This has been done for the Std 6 assessment.

The technical report covers several issues. Questions like the following have been dealt with: what was the overall quality of the tool and how have each of the items functioned with the Bihar student sample. The report looks at information both at test and item level, and provides an opportunity to understand question items that students of all ability levels find to be difficult. In the report, there are clear graphical presentations to depict the status of student ability and tool difficulty levels in comparison to each other.

This kind of detailed and pin pointed feedback about the status of all items for the tools with reference to the ability of the students is valuable to as it gives detailed information not only about the learning level of the students, but also about where are they lagging behind and recommendations on the next steps. Both statistical as well as qualitative analysis methods have been employed to develop this report.

Dr. Rukmini Banerji
Director
ASER Centre, New Delhi

Bihar Elementary School Study

May 2014

Contents

List of Tables

List of Figures

1 | Introduction

Psychometrics

Psychometrics is the field of study concerned with the theory and technique of psychological measurement. One part of the field is concerned with the objective measurement of skills and knowledge, abilities, attitudes, personality traits, and educational achievement. Another part of the field is concerned with statistical research bearing on measurement theory (e.g., item response theory; Classical test theory etc). The latter part of conducting statistical analysis on the item response data has a major focus on four aspects – whether the assessment was reliable, whether the assessment was valid, what are the results or how have the responses been and the effect of the assessment tool on the respondents i.e. whether it was easy or difficult for the respondents to answer.

The information generated by this analysis is divided in two portions – test level and item level. Item here refers to the questions or sub questions each one of which has acted as a stimulus to which there has been some kind of response.

Assessment Analysis

This assessment would be analyzed by employing both Classical test theory and Item response theory. Test level and item level functions will be analyzed qualitatively as well as quantitatively.

Test analysis

Classical test theory is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. Generally speaking, the aim of Classical test theory is to understand and improve the reliability of psychological tests. The basic of CTT is that any observed score X on a test is True score + error in measurement. Reliability and Validity are test level functions in any assessment.

Item analysis

Item analysis is usually done utilizing the Item response theories. There are many kinds of IRT theories as 1, 2, 3 parametric model, Rasch model etc. Each of these models aim towards a common end i.e. checking the performance of item. Each item is analyzed in terms of its difficulty, discrimination and reliability with internal consistency clause.

Reliability of the assessment

In very simple terms, it means that the reliable test items generate similar results without much significant difference of the mean scores every time it is administered on the same population without any intervention made on the learning levels. Tests that have been designed for the same purpose with the same kind of items of same difficulty will also generate similar results. There are various methods to check reliability of an assessment. Usually values above 0.70 are considered to be of good reliability or validity values in educational assessment. It is often seen that reliability and validity of assessments go hand in hand, either both high or both low. There are certain statistical methods which are a measures of both reliability as well as validity.

Parallel forms correlation for reliability and validity

If simple correlation is found between Set A and Set B, these are two Parallel forms made out of the same assessment or test frame. At the same time different people are attempting these two forms that are similar or parallel to each other. Hence it is expected that the results would be similar if the forms are same and people attempting it have similar age and Std based competencies. The correlation between two forms made out of the same frame also is a measure of validity called convergent validity. This depicts that the items in both the assessments have been made out of the same and pre-defined frame which defines the construct. The tests are similar and convergent to each other and they are measuring what they intend to measure.

Split Half correlation

- If the whole set of A is divided into two halves and then we find its correlation it is called split half method to find reliability. Split half can be done by simply dividing the paper into two halves or

selecting even odd items to have a uniform distribution of item difficulty and marks weightage. The average correlation is found using spearman brown formula and cronbach's alpha. Same method is done with set B as well. It basically tells whether the items are consistent with each other and within the defined frame of construct. People attempting these items will perform similarly even if the test items are more or less in number.

Internal consistency based reliability and validity

If we club the total data for set A or total data of set B attempted by different groups of people, we then calculate Point Biserial or internal consistency check. This is a measure of reliability per item and also validity of the construct. It is also often referred to as item total correlation. It basically tells whether performance of one item is consistent with performance of the rest of the items. Are all the items consistent with each other within the defined frame of the construct?

Inter item correlation based reliability and validity

Inter item correlation is also a measure for internal consistency. It is a method to correlate within each set of test, all the items with each other to see whether any one of them is negatively correlated to the other one. A test definitely has a particular frame in which there are types of items which will not always be highly correlated with each other due to certain differences in them but still it is preferred that they must be atleast positively correlated with each other so that performance on one can predict the overall performance on the rest of the items. Inter item correlation is also a measure of both reliability (internal consistency) and validity.

Validity of the assessment

Validity of a test is whether the test is fulfilling its purpose for which it was made. Is the test measuring the construct appropriately or not, how it looks on face, is the content and difficulty level appropriate, has the construct been mapped against suitable curriculum or framework? It is checked by qualitative analysis. Is the test made appropriately according to the table of specification? Is there a Blooms taxonomy match in terms of cognitive levels and their spread item difficulty and its spread across paper? Qualitative analysis is done by subject matter experts who provide normative guidelines (age and Std appropriateness, language, intent etc). Statistical measures for internal consistency are measures for validity are as inter item correlation and item total correlation. Whether all through the test there is consistency in items up to a certain extent would depict in a way that everything has been matched with the pre-defined frame to measure the construct and nothing is deviant or different from the same.

Test level analysis

The test level analysis tells about how the test has performed as a whole. It includes information like the scale statistics, reliability and validity of the tool. Scale statistics has information about the test level mean score and standard deviation. Reliability of the test has been checked utilizing total six methods. Validity of the test has been analyzed utilizing five methods. Inter item correlation and item total correlation methods have been common measures of reliability and validity.

Scale statistics

It can be seen in the following table that the Bihar language assessment for Std 6 consisted of 18 items. Out of 18 items students usually scored on an average 7.28 or could attempt 7.28 items out of 18 items. The standard deviation of which is 4.8.

Table 2.1 Scale statistics- Bihar language assessment Std 6			
Std	Number of Questions	Mean score	Standard Deviation
Std 6	Lang	Lang	Lang
Std 6	18	7.28	4.825

Reliability

Parallel form correlation for reliability achieved a high value of 0.92 indicating high reliability of the test forms A and B. It also indicates that as many times both these tests made under the same frame defining the construct (language) that will be administered on the students of same age, Std and competency the results will be 92% similar until and unless some intervention is given to improve the learning levels. It is also indicative of the fact that the items in the forms A and B are highly similar. They are made out of one frame and are generating same results. The correlation between two forms made out of the same frame also is a measure of validity called convergent validity. This depicts that the items in both the assessments have been made out of the same and pre defined frame which defines the construct. The tests forms are similar and convergent to each other and they are measuring what they intend to measure.

Split half method was also employed by dividing the test into equal halves with even - odd number of items, using simple Spearman brown formula of correlation and the reliability value was found to be 0.79. When cronbach's alpha method was used on the same, the reliability value was 0.88 with a KR 21 (lowest limit of Cronbach's) value of 0.86. Hence it can be concluded that the language test – form A and B were highly reliable assessments for assessing the language competency of Std 6 students in Bihar.

Table 2.2 Reliability coefficients – Bihar language assessment Std 6			
S. No.	Method	Reliability Coefficients[1]	Indication
1.	Parallel form	0.92	Very high
2.	Split half with Spearman Brown formula	0.79	High
3.	Split half with Cronbach's alpha	0.88	Very high
4.	KR21	0.86	Very high

Validity[2]

The validity of a measurement tool in education is considered to be the degree to which the assessment measures what it should measure. In psychometrics, validity has a particular application known as test validity: the degree to which evidence and theory support the interpretations of test scores as entailed by proposed uses of tests. Given below are the different methods used to find validity.

[1]For Reliability analysis Std 6 Language assessment – 95% Confidence interval and 1.68 SEM
[2]Validity analysis includes both qualitative and quantitative analysis. Face validity, content validity and representation validity are qualitative methods whereas convergent validity, criterion validity, inter item correlation and item total correlation for internal consistency are quantitative analysis

Face and content validity

Content validity is a non-statistical type of validity that involves the systematic examination of the test content to determine whether it covers a representative sample of the domain to be measured. Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. For example, a test of math ability to add two numbers should include a range of combinations of digits. Content related evidence typically involves subject matter experts (SME's) evaluating test items against the test specifications. It is also termed face validity as by looking at the face of the test the experts can judge whether it is measuring according to the criterion of the assessable construct or not. A test has content validity built into it by careful selection of which items to include. Items are chosen so that they comply with the test specification which is drawn up by a thorough examination of the subject domain.

The testing assessments were prepared by ASER Centre after analyzing the textbooks and other relevant documents as curriculum and syllabus. All assessments were thoroughly piloted with children of relevant Std in similar type of schools (as our target group). Face validity was achieved by sharing language testing assessments for Std 6 and the respective framework with the experts from SCERT. Their suggestions were incorporated to improve the assessments and contextualized for our target group of children.

For each domain a range of test items in terms of difficulty (easy, medium and difficult) and types (MCQ and non-MCQ) were included by thoroughly mapping the items with the syllabus. For instance, in Std 6 Language testing assessment – vocabulary items from lower Std were included. Some vocabulary items required matching pictures with words and others required choosing the right words to complete sentences and so on. All items were developed at a level below the child's actual Std level in terms of difficulty considering the fact that the student in his current Std has not fully attained the knowledge at this level.

Representation Validity

Representation validity, also known as translation validity, is about the extent to which an abstract theoretical construct can be turned into a specific practical test. The experts are able to review the items and comment on whether the items cover a representative sample of the domain. After detailed mapping of Bihar textbooks and analyzing other documents like curriculum syllabus etc, a detailed assessment framework for language and math was developed for this study. This assessment framework attempts to measure each domain in a number of ways and at varying difficulties. Based on this framework test items were created which were reviewed by subject experts and modified based on the outcomes from pilot field visits. Based on the objective of testing this framework can be used in future to construct more test items measuring performance on specific domains.

Table 2.3 Table of specification- Bihar language assessment Std 6									
Item	Question in Sample 1	Question in Sample 2	Typology	Number of items	Marks	Assessable competency	Difficulty level	Cognitive level	Competencies required
V1	1	1	MCI	1	1	Vocabulary	Easy	Selecting the appropriate word	Reading, comprehending, choosing the correct word
V2	2	4	MCI	1	1	Vocabulary	Easy	Selecting the appropriate word	Reading, comprehending, choosing the correct word
V3	3	2	MCI	1	1	Vocabulary	Easy	Selecting the appropriate word	Reading, comprehending, choosing the correct word
V4	4	3	MCI	1	1	Vocabulary	Difficult	Selecting the appropriate word	Reading, comprehending, choosing the correct word
V5	5	8	MCI	1	1	Vocabulary	Easy	Recall	Reading, comprehending, choosing the correct word
V6	6	6	MCI	1	1	Vocabulary	Difficult	Recall	Reading, comprehending, choosing the correct word
V7	7	5	MCI	1	1	Vocabulary	Easy	Recall	Reading, comprehending, choosing the correct word
V8	8	7	MCI	1	1	Vocabulary	Difficult	Recall	Reading, comprehending, choosing the correct word
Cs1	15	9	MCI	1	1	Reading comprehension	Easy	Direct retrieve	Reading, comprehending, locating a piece of required information and interpreting it, choosing the correct option
Cs2	16	10	MCI	1	1	Reading comprehension	Easy	Indirect retrieve	Reading, comprehending, interpreting information, inferring and choosing the correct option
Cs3	12	11	MCI	1	1	Reading comprehension	Medium	Interpret	Reading, comprehending, interpreting information, evaluating the entire text, presenting ideas by matching previous experience with the current information given in the text, reflecting and writing the response
Cs4	18	12	Constructed response	1	1	Reading comprehension	Difficult	Reflect and evaluate	Reading, comprehending, locating a piece of required information from the text, choosing the correct option
Cs1	9	13	MCI	1	1	Reading comprehension	Easy	Direct retrieve	Reading, comprehending, locating a piece of required information and interpreting it, choosing the correct option
Cs2	10	14	MCI	1	1	Reading comprehension	Easy	Indirect retrieve	Reading, comprehending, locating a piece of required information and interpreting it, choosing the correct option
Cs3	11	15	MCI	1	1	Reading comprehension	Medium	Integrate	Reading, comprehending, locating a piece of required information and interpreting it, choosing the correct option
Cs4	12	16	MCI	1	1	Reading comprehension	Medium	Interpret	Reading, comprehending, locating a piece of required information and interpreting it, choosing the correct option
Cs5	13	17	Constructed response	1	1	Reading comprehension	Difficult	Evaluate	Reading, comprehending, interpreting information, evaluating the entire text, inferring and writing the response
Cs6	14	18	Constructed response	1	1	Reading comprehension	Difficult	Reflect and evaluate	Reading, comprehending, interpreting information, evaluating the entire text, presenting ideas by matching previous experience with the current information given in the text, reflecting and writing the response
Totals			15 MCI and 3 CRI	18	18				

Convergent validity

It involves the correlation between the test and a criterion variable (or variables) taken as representative of the construct. In other words, it compares the test with other measures or outcomes (the criteria) already held to be valid. If both are giving same results either in a good or bad way then it shows validity by being able to match up to the criterion. Criterion validity is further of two kinds – concurrent and predictive. If two assessments measuring same things are administered at the same time to people in same age group and with similar competency then there is a high probability of high positive correlation. This situation is concurrent validity. In Bihar language assessment there is a high correlation between both the forms A and B as seen in the earlier section of the report. This means concurrently two assessments are indicating same kind of results as they have been made out of the same frame for same purpose. Predictive validity is slightly different because some other criterion of the construct which is predictive of the latter is pre utilized to gauge what might happen later on when the original assessment will be applied. Here those students who have performed poorly on the reading assessment or are at the beginner level have also performed poorly on items which require reading, understanding and applying the knowledge to form or pick an appropriate answer. Thus although the reading assessment was used post written assessment yet it can be concluded looking at the following table that those students who are poor in reading levels have performed poorly in written test as well, so reading is a pre requisite of performing well in any written assessment.

Table 2.4 Qualitative analysis for criterion validity Bihar language assessment Std 6								
Assessment Details			**% Children answering question correctly**					**Reading as a Pre Requisite competency for answering all the questions correctly**
Std 6 language pen-paper test		**All children tested**	**Children's performance in written test by reading level**					
QUESTION TYPE	**Variable**		**Std 2 level readers (story)**	**Std 2 level readers (Para)**	**Word level readers**	**Letter level readers**	**Beginner level reader**	**Competencies required to solve these questions correctly**
VOCABULARY		**Q1-Q4 : Choose correct word to complete the given sentence**						
Word usage	v1	51.2	60.9	41	34	28.7	28.1	Reading, comprehending, choosing the correct word
	v2	57.4	67.8	46.4	41	32	27.1	Reading, comprehending, choosing the correct word
	v3	57.9	67.2	50	42	33.3	29.7	Reading, comprehending, choosing the correct word
	v4	43.2	50.4	36.3	29.4	26.7	21.3	Reading, comprehending, choosing the correct word
VOCABULARY		**Q5-Q8 : Word usage & vocabulary**						
Find synonym word	v5	50.9	60.6	41.6	32.5	27.7	24.4	Reading, comprehending, choosing the correct word
Find antonym word	v6	32.3	36.9	28.7	25.6	18.6	18.3	Reading, comprehending, choosing the correct word
	v7	47.9	56.6	39.5	32.7	26.2	25.8	Reading, comprehending, choosing the correct word
Find correct meaning of proverb (muhavara)	v8	52.5	61.3	46.5	34.8	29.6	25.8	Reading, comprehending, choosing the correct word
READING & COMPREHENSION		**The next set of questions are for reading & comprehension.**						
Direct fact retrieval	cs1	61	73.6	49.8	36.3	30.7	26.2	Reading, comprehending, locating a piece of required information from the text, choosing the correct option
Indirect fact retrieval	cs2	40.4	48.3	33.4	25.6	21.3	16.7	Reading, comprehending, locating a piece of required information and interpreting it, choosing the correct option
Interpret	cs3	41	49.8	30.9	24.6	21.9	19	Reading, comprehending, linking information provided at different locations of the text, choosing the correct option
Reflect (open ended question)	cs4	14.7	20.3	7.3	4.2	3.9	4.4	Reading, comprehending, interpreting information, evaluating the entire text, presenting ideas by matching previous experience with the current information given in the text, reflecting and writing the response
Direct fact retrieval	cw1	48.4	58.2	38.7	30.6	24.1	23.2	Reading, comprehending, locating a piece of required information from the text, choosing the correct option
Indirect fact retreival	cw2	37.3	44.8	31	23.3	17.9	15.1	Reading, comprehending, locating a piece of required information and interpreting it, choosing the correct option
Integrate	cw3	34.8	41	28.5	23.2	21.3	15.3	Reading, comprehending, linking information provided at different locations of the text, choosing the correct option
Interpret	cw4	18.3	20.6	15	16.4	12.4	10.7	Reading, comprehending, interpreting information, inferring and choosing the correct option
Analyze (open ended question)	cw5	23.6	32.4	12.5	7.4	6.6	5.6	Reading, comprehending, interpreting information, evaluating the entire text, inferring and writing the response
Reflect (open ended question)	cw6	14.2	20.2	6	2.9	3.4	0.9	Reading, comprehending, interpreting information, evaluating the entire text, presenting ideas by matching previous experience with the current information given in the text, reflecting and writing the response
Total children tested	**17648**							

Inter item correlation[3]

Inter item correlation shows how well do the items internally correlate and how consistent they are in predicting each other's performance. The items which have a negative correlation with most of the other items show that there is some problem of internal consistency. The possible reasons could be that the content is different or construct has too many contexts or there is a language inconsistency or difficulty level inconsistency, and items are assessing completely different things. A negative correlation is not good as it means performance on one item would be different from another item and hence they are not predictive. A zero correlation is still better as it means there is no correlation between one item with the other item. Positive correlation is ideal. The more the number of negative correlation the worse the item becomes in internal consistency issues and predictability. This is a measure of internal consistency of the test. Internal consistency also means reliability as well as validity. In this study we find that none of the items negatively correlate with each other. On finding the average inter item correlation the value indicates 0.28 which is in positive direction and a good indication of none the items being completely different from the whole test or the frame of the construct defined.

Table 2.5 Inter item correlation for Bihar language assessment Std 6																		
Items	V1	V2	V3	V4	V5	V6	V7	V8	Cs1	Cs2	Cs3	Cs4	Cw1	Cw2	Cw3	Cw4	Cw5	Cw6	Total score
V1	1	0.42	0.45	0.39	0.41	0.25	0.42	0.36	0.4	0.26	0.32	0.2	0.36	0.29	0.27	0.16	0.27	0.24	0.64
V2	0.42	1	0.44	0.38	0.41	0.24	0.39	0.39	0.37	0.27	0.33	0.21	0.33	0.31	0.25	0.14	0.28	0.23	0.63
V3	0.45	0.44	1	0.41	0.4	0.27	0.4	0.36	0.38	0.27	0.33	0.2	0.32	0.29	0.26	0.16	0.26	0.23	0.64
V4	0.39	0.38	0.41	1	0.38	0.28	0.38	0.33	0.3	0.24	0.28	0.17	0.29	0.28	0.24	0.17	0.25	0.22	0.59
V5	0.41	0.41	0.4	0.38	1	0.31	0.43	0.42	0.38	0.27	0.33	0.2	0.34	0.32	0.27	0.17	0.28	0.24	0.65
V6	0.25	0.24	0.27	0.28	0.31	1	0.3	0.28	0.2	0.16	0.19	0.09	0.2	0.24	0.18	0.15	0.14	0.13	0.46
V7	0.42	0.39	0.4	0.38	0.43	0.3	1	0.45	0.37	0.28	0.33	0.22	0.36	0.31	0.28	0.18	0.29	0.26	0.66
V8	0.36	0.39	0.36	0.33	0.42	0.28	0.45	1	0.36	0.25	0.29	0.2	0.32	0.3	0.25	0.15	0.26	0.23	0.62
Cs1	0.4	0.37	0.38	0.3	0.38	0.2	0.37	0.36	1	0.38	0.4	0.22	0.41	0.32	0.3	0.17	0.29	0.23	0.64
Cs2	0.26	0.27	0.27	0.24	0.27	0.16	0.28	0.25	0.38	1	0.39	0.19	0.28	0.25	0.25	0.19	0.23	0.19	0.53
Cs3	0.32	0.33	0.33	0.28	0.33	0.19	0.33	0.29	0.4	0.39	1	0.27	0.35	0.33	0.32	0.23	0.31	0.29	0.61
Cs4	0.2	0.21	0.2	0.17	0.2	0.09	0.22	0.2	0.22	0.19	0.27	1	0.21	0.18	0.21	0.09	0.45	0.46	0.44
Cw1	0.36	0.33	0.32	0.29	0.34	0.2	0.36	0.32	0.41	0.28	0.35	0.21	1	0.34	0.34	0.2	0.29	0.24	0.61
Cw2	0.29	0.31	0.29	0.28	0.32	0.24	0.31	0.3	0.32	0.25	0.33	0.18	0.34	1	0.32	0.18	0.27	0.24	0.57
Cw3	0.27	0.25	0.26	0.24	0.27	0.18	0.28	0.25	0.3	0.25	0.32	0.21	0.34	0.32	1	0.23	0.26	0.25	0.54
Cw4	0.16	0.14	0.16	0.17	0.17	0.15	0.18	0.15	0.17	0.19	0.23	0.09	0.2	0.18	0.23	1	0.12	0.14	0.36
Cw5	0.27	0.28	0.26	0.25	0.28	0.14	0.29	0.26	0.29	0.23	0.31	0.45	0.29	0.27	0.26	0.12	1	0.55	0.55
Cw6	0.24	0.23	0.23	0.22	0.24	0.13	0.26	0.23	0.23	0.19	0.29	0.46	0.24	0.24	0.25	0.14	0.55	1	0.49
Total score	0.64	0.63	0.64	0.59	0.65	0.46	0.66	0.62	0.64	0.53	0.61	0.44	0.61	0.57	0.54	0.36	0.55	0.49	1

[3]Inter item correlation shows both reliability and validity. It is a measure of internal consistency. It shows that scores in one item are matching and predicting with the remaining items of the assessment as well as whether all items are in sync with each other and measuring what they intend to do as a whole within the defined construct.

Item total correlation[4]

This correlation is the regressive correlation done between the scoring data of each of the items situated in different columns and the total raw score in the last column for all the students on all the items. In this concept a higher correlation means that performance of all the students on one item is predictive of their performance on remaining items or the test. A zero correlation means performance on that item does not predict or impact performance on the remaining items or the test. This is again a measure of internal consistency in the test. Internal consistency also means reliability as well as validity. Here we see that all items are positively correlated with the total and some items have been highlighted because the correlation was less than .50 but even on doing further analysis it was found that removing any one of that item would reduce the quality of the test by lowering the consistency. Hence all these items together make a valid assessment.

Table 2.6 Item total correlation- Bihar language assessment Std 6		
Item	Item total correlation	Corrected item total correlation
V1	0.63	0.57
V2	0.62	0.56
V3	0.63	0.56
V4	0.59	0.52
V5	0.64	0.58
V6	0.45	0.37
V7	0.65	0.59
V8	0.61	0.54
Cs1	0.63	0.57
Cs2	0.52	0.45
Cs3	0.61	0.55
Cs4	0.47	0.38
Cw1	0.61	0.54
Cw2	0.56	0.49
Cw3	0.54	0.46
Cw4	0.37	0.28
Cw5	0.57	0.49
Cw6	0.53	0.45

Item level analysis

This includes item information, item person map, item difficulty, Distracter analysis of the MCI, Item discrimination, Item characteristic curves for the items with strange performance

[4]Item total correlation is a measure of both reliability and validity. It shows whether there is a similar trend and positive correlation between responses of one item with the total result of the individual or the group. Any item would be predictor for the rest of the items thus. It also shows whether all items are performing in sync with the rest of the items as a whole and falls within the defined limit of the measurable construct or not.

Item difficulty:

The term item difficulty or p value expresses the proportion or percentage of students who answered the item correctly. Item difficulty can range from 0.0 (none of the students answered the item correctly) to 1.0 (all of the students answered the item correctly). Item difficulty is also termed as Item facility.

Figure 2.1 Understanding item difficulty											
0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Nobody is able to solve this item correctly		20%	30%	40%	50 percent could solve this item	60%	70%	80%	Everybody is able to solve this item correctly	
Very Difficult item		Difficult	Difficult	Medium	Medium	Medium	Easy	Easy	Very easy item	

The normal distribution curve has a spread of population of all kinds i.e. less able to average to more able. As per the Rasch model of item analysis if the item difficulty and the person ability are calibrated on the same scale it would appear as the figure given below.

Figure 2.2 Understanding item difficulty and person ability																					
Location of Less able group										Average ability group	Location of More able group										
-1.0	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	+0.8	+0.9	+1.0	
Negative sided S.D. or standard deviation from mean										MEAN	Positive sided S.D. or standard deviation from mean										
Location of Easier items										Medium items	Location of Difficult items										

Table 2.7 Item difficulty set wise- Bihar language assessment Std 6

Set A	p value 1	Set B	p value 2	Difficulty Level
V1_1/1	0.43	V1_2/1	0.6	Both Medium
V2_1/2	0.63	V2_2/4	0.52	Both Medium
V3_1/3	0.59	V3_2/2	0.57	Both Medium
V4_1/4	0.44	V4_2/3	0.42	Both Medium
V5_1/5	0.52	V5_2/8	0.49	Both Medium
V6_1/6	0.35	V6_2/6	0.29	Both Difficult
V7_1/7	0.47	V7_2/5	0.49	Both Medium
V8_1/8	0.54	V8_2/7	0.52	Both Medium
Cs1_1/15	0.56	Cs1_2/9	0.66	Both Medium
Cs2_1/16	0.39	Cs2_2/10	0.42	A slightly difficult then B
Cs3_1/17	0.4	Cs3_2/11	0.42	Both Medium
Cs4_1/18	0.14	Cs4_2/12	0.15	Both Difficult
Cw1_1/9	0.47	Cw1_2/13	0.5	Both Medium
Cw2_1/10	0.41	Cw2_2/14	0.34	B slightly difficult then A
Cw3_1/11	0.35	Cw3_2/15	0.35	Both Difficult
Cw4_1/12	0.18	Cw4_2/16	0.19	Both Difficult
Cw5_1/13	0.24	Cw5_2/17	0.24	Both Difficult
Cw6_1/14	0.15	Cw6_2/18	0.14	Both Difficult
Total Difficult items	7	Total Difficult items	7	Equal
Total Medium items	11	Total Medium items	11	Equal
Total Easy items	0	Total Easy items	0	Equal
Total Items	**18**	**Total Items**	**18**	**Equal**

Important :

The items which have the biggest p value in Set A and B are marked in Yellow which denotes that these were attempted by large number of people and were comparatively easy.

The items which have the smallest p value in Set A and B are marked in Red which denotes that these were attempted by least number of people and were comparatively difficult.

Index

p-values	Indicator	Difficulty Level
0.0 to 0.39		Difficult
0.4 to 0.69		Medium
0.7 to 1.0		Easy

Item discrimination

What is Findlay's Index and Item discrimination

We want items on our tests that inform us about student learning. We want items that differentiate or discriminate between those who score high and low on the test. We want this index to be easy to compute and easy to understand. The Findlay Index is the easiest to use. There are three kinds of items.

- Over Discriminating items – The poor students do worse and good students do more than expected.

- Under Discriminating items – The good students perform worse or less than expectation and the poor students surprisingly do better. These are invalid items.

- Discriminating item – Students perform according to their caliber and probability* to get the answer correct.

Findlay Index formula:

Rank order students in descending order based on total test score.

Split these students into three groups:

NRU = Number of students getting the item correct in the upper band of the test

NRL = Number of students getting the item correct in the lower band of the test

NU = Number of total students in the upper band of the test

(Ignore the middle band of students)

$$\text{Findlay Index} = \frac{\text{No. correct in Upper band} - \text{No. correct in Lower band}}{\text{No. of students in Upper band}}$$

The number in the upper and lower bands must be the same.

Figure 2.3 Understanding Item discrimination			
DISCRIMINATION TYPE	OBSERVED RESPONSE	EXPECTED RESPONSE	ACCEPTANCE
OVER DISCRIMINATING	GOOD STUDENTS DID EVEN BETTER THAN EXPECTED AND POOR STUDENTS DID WORSE THAN EXPECTED	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	x
NORMAL DISCRIMINATING	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	✓
UNDER DISCRIMINATING	GOOD STUDENTS DID WORST THAN EXPECTED AND POOR STUDENTS DID MUCH BETTER THAN EXPECTED	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	x

Table 2.8 A comprehensive table to understand the item function - competency wise difficulty and discrimination- Bihar language assessment Std 6				
Item	Difficulty	Interpretation	Discrimination	Interpretation
v1	0.51	Medium	0.57	Students are performing as per their ability
v2	0.51	Medium	0.56	Students are performing as per their ability
v3	0.51	Medium	0.57	Students are performing as per their ability
v4	0.43	Medium	0.52	Students are performing as per their ability
v5	0.5	Medium	0.58	Students are performing as per their ability
v6	0.32	Difficult	0.37	Students are not performing as per their ability
v7	0.48	Medium	0.59	Students are performing as per their ability
v8	0.52	Medium	0.54	Students are performing as per their ability
cs1	0.6	Most correctly attempted Medium	0.57	Students are performing as per their ability
cs2	0.4	Medium	0.44	Students are not performing as per their ability
cs3	0.4	Medium	0.54	Students are performing as per their ability
cs4	0.14	Most Difficult	0.37	Students are not performing as per their ability
cw1	0.48	Medium	0.53	Students are performing as per their ability
cw2	0.37	Difficult	0.49	Students are not performing as per their ability
cw3	0.34	Difficult	0.45	Students are not performing as per their ability
cw4	0.18	Difficult	0.28	Students are not performing as per their ability
cw5	0.23	Difficult	0.47	Students are not performing as per their ability
cw6	0.14	Most Difficult	0.43	Students are not performing as per their ability

Table 2.9 Item location and corresponding performance- Bihar language assessment Std 6			
Item	Location	Interpretation as per location on RASCH scale	Percentage of people attempting correctly
cs1	-1.37	Easiest Item	60%
v3	-1.18	Easy	51%
v2	-1.14	Easy	51%
v8	-0.84	Easy	52%
v1	-0.75	Easy	51%
v5	-0.72	Easy	50%
cw1	-0.57	Easy	48%
v7	-0.54	Easy	48%
v4	-0.24	Easy	43%
cs3	-0.09	Medium	40%
cs2	-0.06	Medium	40%
cw2	0.13	Medium	37%
cw3	0.31	Medium	34%
v6	0.47	Medium	32%
cw5	1.10	Difficult	23%
cw4	1.57	Difficult	18%
cs4	1.91	Difficult	14%
cw6	1.99	Most Difficult	14%

Distracter Analysis

The N represents those children who have marked only one option out of the possible four MCI answers a b c d i.e. those children who have not marked option don't know or left the question blank or marked outside the box anywhere or marked multiple ticks. Open ended Question have been ignored to calculate this N. Only MCI were looked at this N. The reason for collecting and analyzing this N is based on the fact that a good MCI gives 25% probability to all the four options to be selected. Thus in distracter analysis if any option gets more than 25% opportunity the option needs to be seen whether it is a strong distracter or does it reveal some typical misconception of the student.

Table 2.10 Distracter analysis of MCI- Bihar language assessment Std 6 – selected sample															
N = 7213 Distracter Analysis (Only for the students who understand concept of MCI and mark on given options)															
Set A	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15
Question	1	2	3	4	5	6	7	8	9	10	11	12	15	16	17
Option A	13.1	83.1	6.8	15.7	8.1	14.7	4.3	72.4	14.7	13	16.2	14.1	8.4	28	10.3
Option B	11.6	6.6	79.5	9.7	8.5	50.5	13.4	5.2	5.2	59.5	24	30.3	5.7	7.5	60.7
Option C	12.5	7.5	8.7	9.3	73	25	17.6	13.3	15.5	13.2	49.7	29.2	78.9	6.8	19.1
Option D	62.7	2.8	5.1	65.3	10.4	9.9	64.7	9.1	64.6	14.4	10.1	26.4	7	57.6	9.9
Set B	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14	Item15
Question	1	2	3	4	5	6	7	8	9	10	11	13	14	15	16
Option A	11.7	9.5	8.7	72.8	5.5	13.8	7.2	9.9	5.7	27	9.4	12.5	13.5	15.9	33.5
Option B	80.2	9.4	24.7	5.4	68	29	71.1	6.9	86	8.1	20.1	69	20.9	51.9	31.2
Option C	3.4	76.8	59.1	9.4	12.7	15	13	12	2.6	59.6	60.5	15.5	13.3	21.8	27.8
Option D	4.6	4.3	7.4	12.4	13.8	42.2	8.7	71.3	5.7	5.3	10	3.1	52.3	10.4	7.5

Index for Options	
Equal or Above .50	Very Strong Distracter
Equal or Above .35	Strong Distracter
Equal or Above .25	Distracter
A/B/C/D	Answer key

Index for Items in terms of difficulty	
	Difficult
	Medium
	Easy

Analysis of errors in MCI

These comparative histograms are showing the movement of number of students and their mean scores in each group as per the criterion. The criterion are : All students, students who have not done errors in picking options a-b-c-d with codes 1-2-3-4 correctly in MCI, students who have done errors minimum 5 to maximum. The errors can be defined and coded as - 0- blank, 5-dont know, 8- Multiple ticks and 9-marked outside the box. The points to be observed in these graphs are the red arrows and blue dots. This would show where the maximum number of students in each category are and what their mean score is. If all graphs are observed together then the shift of mean can easily be identified.

Figure 2.4 Comparative Analysis Of All Students With Students Who Did Not Do Any Error, Students Who Did Less Than Or More Than 5 Errors In Attempting MCI
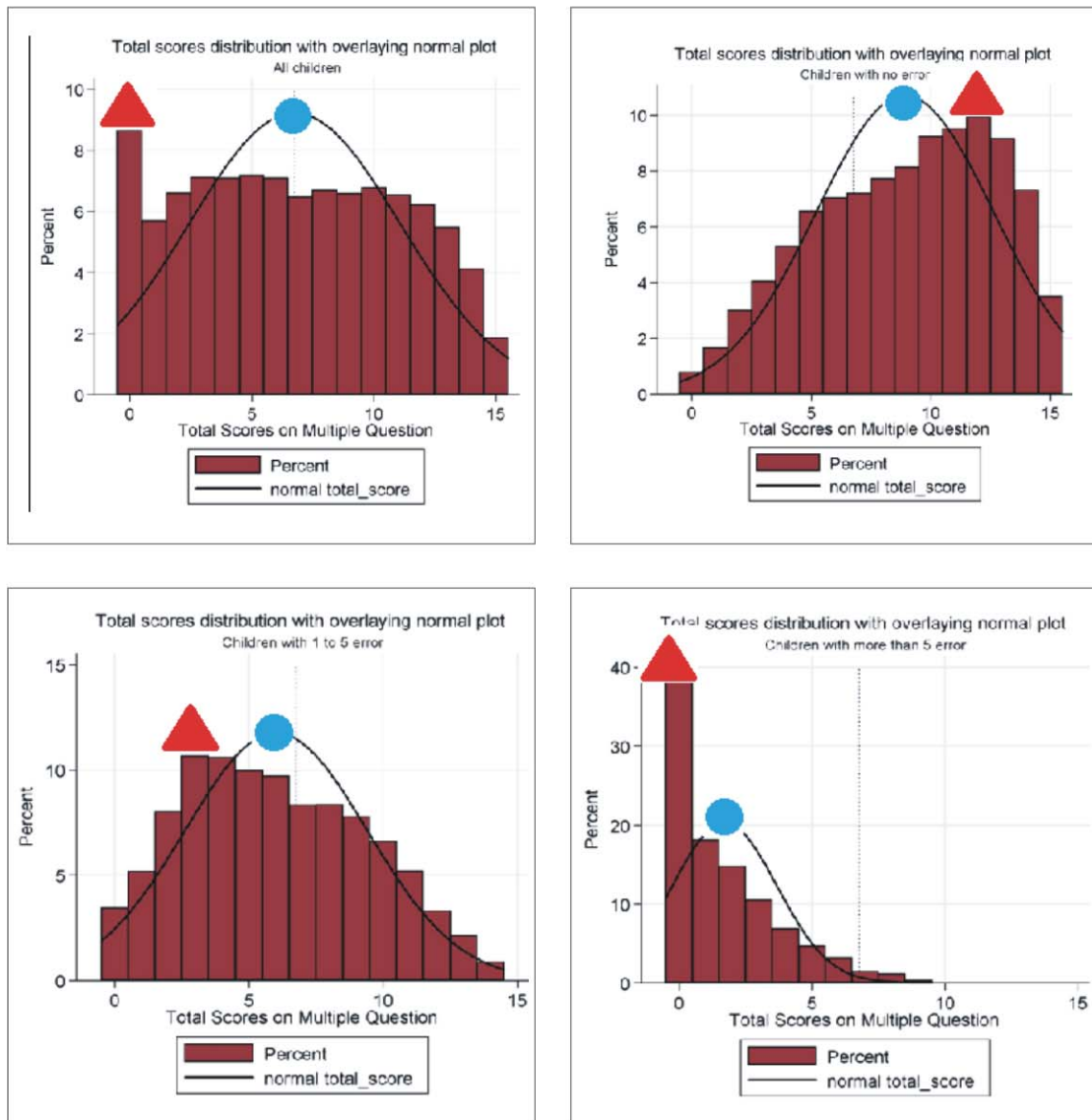
▲ This red arrow denotes the peak of bars which means maximum number of people are located here

● This blue dot shows the mean score of each of these groups of students on all multiple choice items

Item Person Map

The given graph is known as the Item Person Map. It depicts the spread of item and persons on a common item-person scale. The mean location of item is 0.00 which means that average student here is located towards the negative side of the items where he or she can attempt the less difficult items with ease. The kurtosis for person map (peak of curve) is -1.01 and it is skewed (slope) at 0.21.

The distribution of the persons in the given figure reveal that more number of students are on the left side (less able than average) of the scale in comparison to the number of able students are towards the right side (more able than average). The difficulty levels of almost 8 to 12 of the items are matching with ability of the students. But these items are all located close to the mean and not of very easy level for this group of people. Remaining 6 items are difficult even for the average student located closer to the mean.

Figure 2.5 Item person map Bihar language assessment Std 6
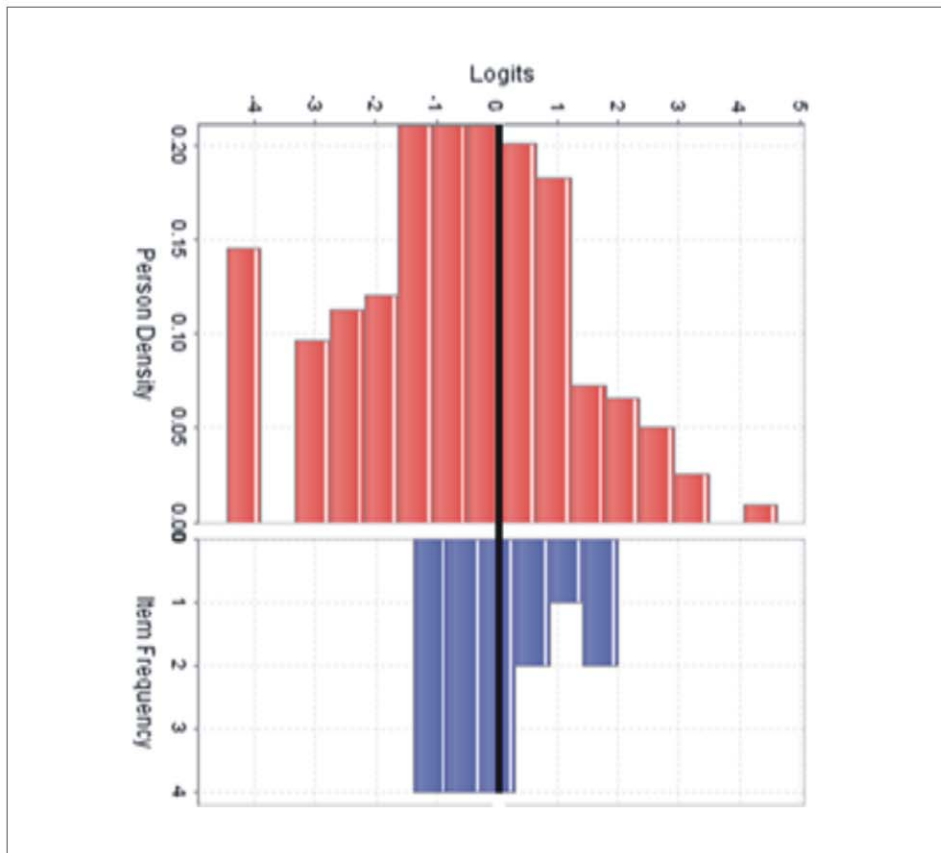
There was a small group of people who are above the average group and away from the mean on the right side. All items are easy for them. There is a group who are located on the extreme negative side of the mean and found all the items difficult then their ability level. Remaining people are close to the mean on both sides and they found items of both categories medium to difficult, but no item was easy for them overall.

Percentile Bands

If we see the student's performance with reference to what should have been their mean scores we find that very few students are in the highest percentile band. Where as if we see student's performance with reference to what has been their mean scores we find that there are almost equally distributed across all percentile bands. Also the last percentile band has a wide range of scores in comparison to the expected scores in the last percentile band. The shift is more towards the negative side of the mean in the real observed case. The mean in each percentile band and the mean within each one of these bands reflect the same. This overall reflects that the students have performed poorly in comparison to what was expected.

Table 2.11 Percentile bands for students who attempted Bihar language assessment Std 6					
PERCENTAGE OF STUDENTS	27.02%	26.58%	23.35%	23.05%	Notice the increase in Percentage of students in the most able group when there is a broad range of scores to be achieved
PERCENTILE BAND	25TH	50TH	75TH	100TH	
RANGE OF PERCENTAGE OF MARKS	UPTO 25%	UPTO 50%	UPTO 75%	UPTO 100%	
ACTUAL MEAN	1.4	5.5	9.5	14	
MEAN SCORE WITHIN	0 to 3	4 to 7	8 to 11	12 to 18	
ACTUAL TEST MEAN SCORE		7.28			Notice the shifted mean in Actual test mean score. Students on an average should have got the expected mean scores.
TEST SCORES	0 to 18				
EXPECTED TEST MEAN SCORE		9			
MEAN SCORE WITH IN	0 TO 4.5	4.6 TO 9	9.1 TO 13.5	13.6 TO 18	
EXPECTED MEAN	2	6.9	11.4	15.3	
RANGE OF PERCENTAGE OF MARKS	UPTO 25%	UPTO 50%	UPTO 75%	UPTO 100%	
PERCENTILE BAND	25TH	50TH	75TH	100TH	Notice the small Percentage of most able students who have performed as per expectation even when there is a narrow range of good scores to be achieved
PERCENTAGE OF STUDENTS	33.80%	31.70%	21.80%	12.70%	

Item benchmarking

All the students are put in specific percentile bands, then each of the item is seen in terms of the number of correct responses in each band. Those items in a particular percentile band which have percentage of students greater than or equal to 50% are suitable for such ability groups. Here we find that 3 items can't be bench marked which means 50% of the most able students also could not attempt these items correctly, 4 items are benchmarked at the most able level which means 50% of the most able students could attempt these items correctly, 8 items are benchmarked at the above average level, 3 items were benchmarked at the average level and none of the items are benchmarked at the low ability level. This shows that the students located on the extreme negative end of the mean and those who are in the 25th percentile band did not find any item suitable for them in terms of getting it attempted correctly.

Ability groups	Low ability	Average	Above Average	Most able	Benchmarked at which ability group
Table 2.12 Item benchmarking: Bihar language assessment Std 6					
Items	25th Percentile	50th Percentile	75th Percentile	100th Percentile	
V1	%	%	%	%	
0	90.2	59.7	24.5	5.9	
1	9.8	40.3	75.5	94.1	Above Average
V2	%	%	%	%	
0	84.5	50.8	18.1	3.2	
1	15.5	49.2	81.9	96.8	Average
V3	%	%	%	%	
0	86.3	47.8	17.3	3.5	
1	13.7	52.2	82.7	96.5	Average
V4	%	%	%	%	
0	92.5	67.5	40.1	12	
1	7.5	32.5	59.9	88	Above Average
V5	%	%	%	%	
0	91	60.3	25.1	5.5	
1	9	39.7	74.9	94.5	Above Average
V6	%	%	%	%	
0	92.1	76.2	58.5	32.1	
1	7.9	23.8	41.5	67.9	Most Able
V7	%	%	%	%	
0	92.4	65.9	28	6.2	
1	7.6	34.1	72	93.8	Above Average
V8	%	%	%	%	
0	86.4	56.6	27.1	4.5	
1	13.6	43.4	72.9	95.5	Above Average
Cs1	%	%	%	%	
0	85.9	39.9	14.8	3.3	
1	14.1	60.1	85.2	96.7	Average
Cs2	%	%	%	%	
0	92	66.6	46.2	20.9	
1	8	33.4	53.8	79.1	Above Average
Cs3	%	%	%	%	
0	93.9	70.9	44.9	11.3	
1	6.1	29.1	55.1	88.7	Above Average
Cs4	%	%	%	%	
0	97.7	93.3	82.8	58.1	
1	2.3	6.7	17.2	41.9	Can't be benchmarked
Cw1	%	%	%	%	
0	90.4	61	31.9	8.7	
1	9.6	39	68.1	91.3	Above Average
Cw2	%	%	%	%	
0	93	74.5	51.3	17.1	
1	7	25.5	48.7	82.9	Most Able
Cw3	%	%	%	%	
0	91.9	76.3	58.8	19.9	
1	8.1	23.7	41.2	80.1	Most Able
Cw4	%	%	%	%	
0	95.9	86.9	81.1	54.7	
1	4.1	13.1	18.9	45.3	Can't be benchmarked
Cw5	%	%	%	%	
0	95.3	87.5	69.6	41.7	
1	4.7	12.5	30.4	58.3	Most Able
Cw6	%	%	%	%	
0	98.1	94.9	83.4	58	
1	1.9	5.1	16.6	42	Can't be benchmarked

Figure 2.6 Pie chart representation of all language items 6
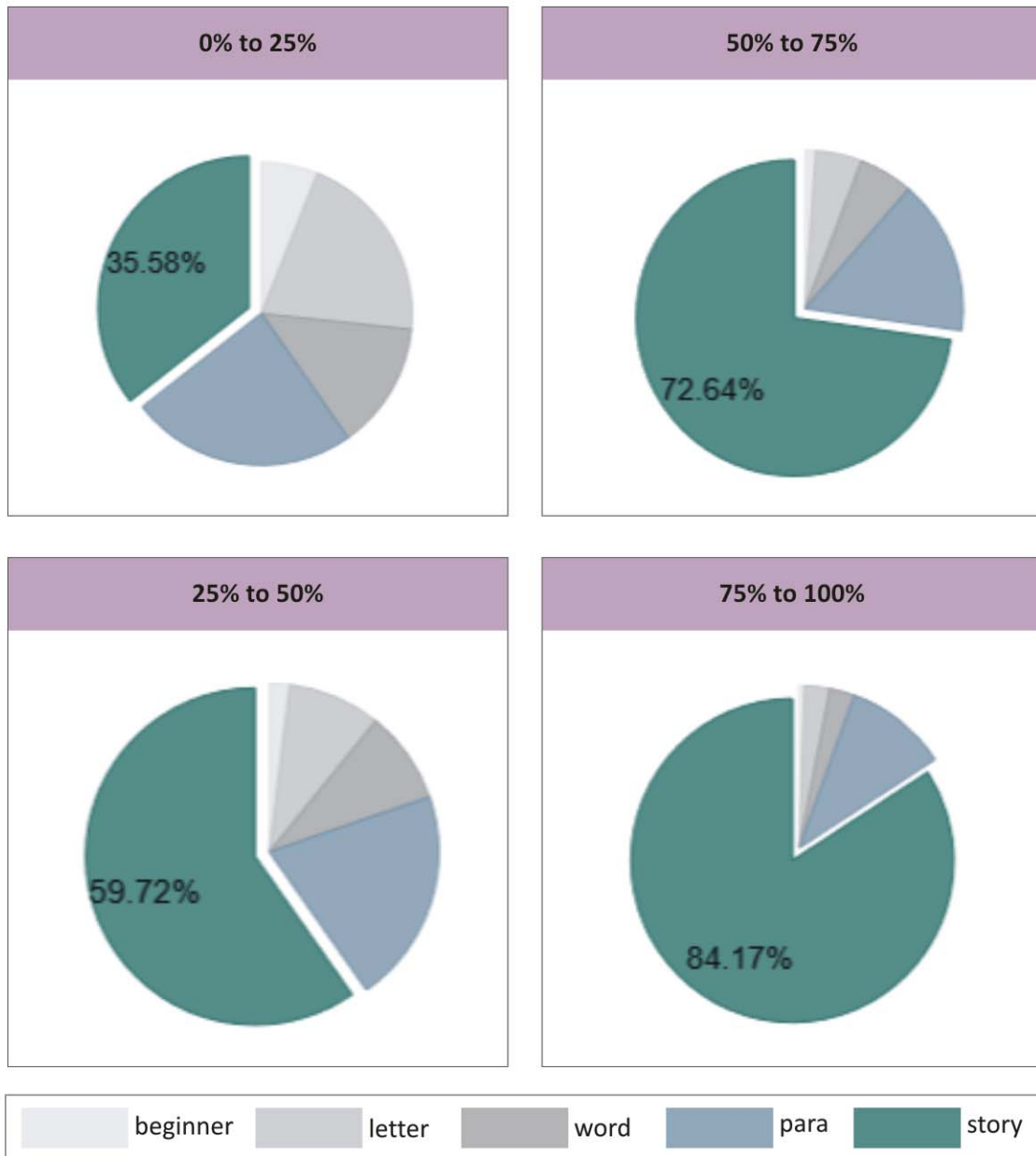
0% to 25%

35.58%

50% to 75%

72.64%

25% to 50%

59.72%

75% to 100%

84.17%

beginner   letter   word   para   story

Key Observations

Observe the increasing number of story level students in each of the percentile band from 25th to 100th

Observe the beginner level students who do not know reading. Their number decreases as the percentile band progresses

Observe the letter, word and para level students. Even their numbers decreases as the percentile band progresses Maximum number of story level students are in the top most percentile band

Conclusive remark (Std 6 Bihar Language Assessment)

The students could on an average achieve less than the ideal mean score. Their ability clusters more on the negative side of the mean and their probability to get full scores decreases with their lowering ability. A very small group of persons who are located on the extreme positive side of the mean have a higher probability to achieve full scores where as the group of people who are located on the extreme negative side of the mean do not have significant or visible probability to attempt the items correctly. The gap between the most able group and the average people is wider than the average and least able group. The test items are pitched at a lower Std level for the students yet they do not find any item easy, every item for them is medium to difficult.

Figure 2.7 Person plot Bihar language assessment Std 6
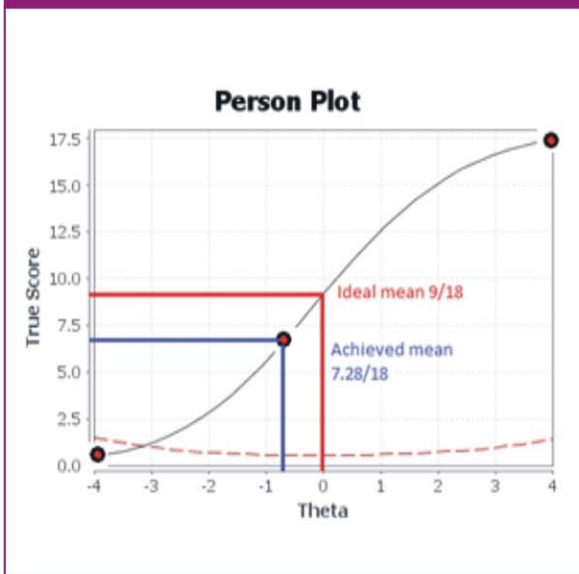
Figure 2.8 Item characteristic curves – the most difficult item Bihar language assessment Std 6
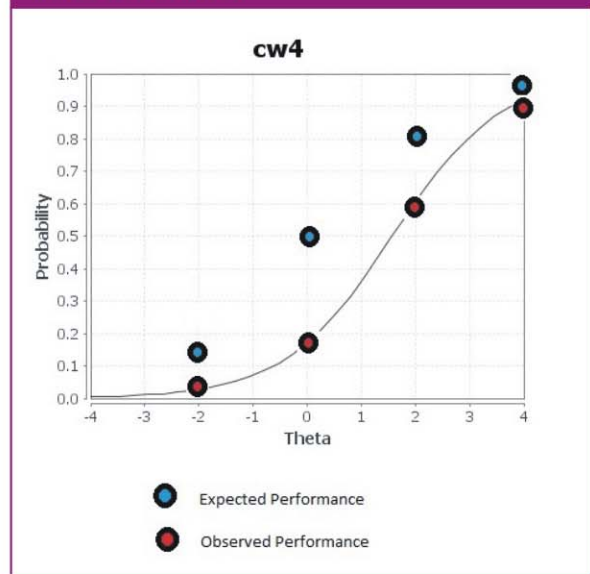
Figure 2.9 Item characteristic curves - the most weakly discriminating item Bihar language assessment Std 6
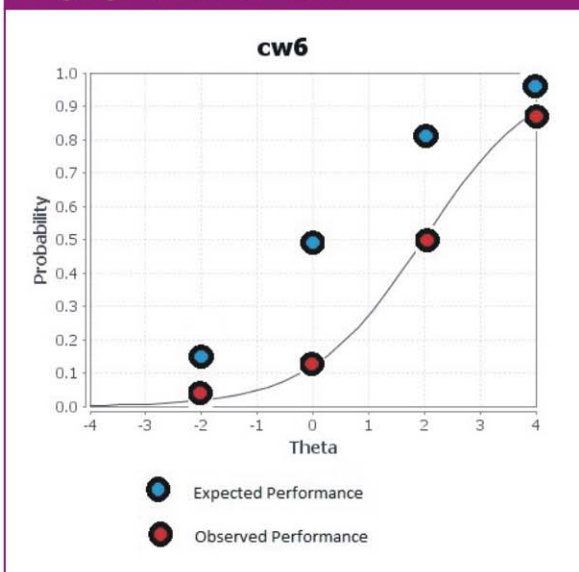
Test level analysis

The test level analysis tells about how the test has performed as a whole. It includes information like the scale statistics, reliability and validity of the tool. Scale statistics has information about the test level mean score and standard deviation. Reliability of the test has been checked utilizing total six methods. Validity of the test has been analyzed utilizing five methods. Inter item correlation and item total correlation methods have been common measures of reliability and validity.

Scale statistics

It can be seen in the following table that the Bihar math assessment for Std 6 consisted of 41 items. Out of 41 items student scored on an average 16.43 marks or could attempt 16.43 items out of 41 items. The standard deviation of which is high 9.03.

Table 3.1 Scale statistics- Bihar mathematics assessment Std 6			
Std	**Number of Questions**	**Mean score**	**Standard Deviation**
Std 6	Math	Math	Math
Std 6	41	16.43	9.03

Reliability

Parallel form correlation for reliability achieved a coefficient value of 0.97 which is a very high value indicating high reliability of the test forms A and B. It indicates that as many times both these tests are administered under the same frame defining the construct (Math) on the students of same age, Std and competency the results will be 97% similar. Until and unless some intervention is given to improve the learning levels. It is also indicative of the fact that the items in the forms A and B are highly similar. They are made out of one frame and are generating same results. The correlation between two forms made out of the same frame also is a measure of validity called convergent validity. This depicts that the items in both the assessments have been made out of the same and pre defined frame which defines the construct. The tests are similar and convergent to each other and they are measuring what they intend to measure.

Split half method was employed by dividing the test into equal halves with even - odd number of items, using simple Spearman brown formula of correlation and the reliability coefficient value was 0.89. When cronbach's alpha method was used on the same the reliability value was 0.92 with a KR 21 (lowest limit of Cronbach's) value of 0.90.

Hence it can be concluded that the Math test – form A and B were highly reliable assessments for assessing the math competency of Std 6 students in Bihar.

Table 3.2 Reliability coefficient- Bihar mathematics assessment Std 6			
S. No.	**Method**	**Reliability Coefficients[5]**	**Indication**
1.	Parallel form	0.92	Very high
2.	Split half with Spearman Brown formula	0.79	High
3.	Split half with Cronbach's alpha	0.88	Very high
4.	KR21	0.86	Very high

Validity[2]

The validity of a measurement tool in education is considered to be the degree to which the assessment measures what it claims to measure. In psychometrics, validity has a particular application known as test validity: the degree to which evidence and theory support the interpretations of test scores as entailed by proposed uses of tests. Given below are the different methods used to find validity.

---

[5] For Reliability analysis Std 6 Language assessment – 95% Confidence interval and 1.68 SEM
[6] Validity analysis includes both qualitative and quantitative analysis. Face validity, content validity and representation validity are qualitative methods whereas convergent validity, criterion validity, inter item correlation and item total correlation for internal consistency are quantitative analysis

Face and content validity

Content validity is a non-statistical type of validity that involves the systematic examination of the test content to determine whether it covers a representative sample of the domain to be measured. Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. For example, a test of math ability to add two numbers should include a range of combinations of digits. Content related evidence typically involves subject matter experts (SME's) evaluating test items against the test specifications. It is also termed face validity as by looking at the face of the test the experts can judge whether it is measuring according to the criterion of the assessable construct or not. A test has content validity built into it by careful selection of which items to include. Items are chosen so that they comply with the test specification which is drawn up by a thorough examination of the subject domain.

The testing assessments were prepared by ASER Centre after analyzing the textbooks and other relevant documents as curriculum and syllabus. All assessments were thoroughly piloted with children of relevant Std in similar type of schools (as our target group). Face validity was achieved by sharing math testing assessments and the respective frameworks with the experts from SCERT. Their suggestions were incorporated to improve the assessments and contextualize them better for our target group of children.

For each domain a range of test items in terms of difficulty (easy, medium and difficult) and types (MCQ and non-MCQ) were included by thoroughly mapping the items with the syllabus. All items have been tried to be developed at a level below the child's actual Std level in terms of difficulty considering the fact that the student in his current Std and has not fully attained the knowledge at this level.

Representation Validity

Representation validity, also known as translation validity, is about the extent to which an abstract theoretical construct can be turned into a specific practical test. The experts are able to review the items and comment on whether the items cover a representative sample of the domain. After detailed mapping of Bihar textbooks and analyzing other documents like curriculum syllabus etc, a detailed assessment framework for math was developed for this study. This assessment framework attempts to measure each domain in a number of ways and at varying difficulties. Based on this framework, test items were created which were reviewed by subject experts and modified based on the outcomes from pilot field visits. Based on the objective of testing this framework can be used in future to construct more test items measuring performance on specific domains.

Table 3.3 Table of specification- Bihar mathematics assessment Std 6						
Std 6	*Please note – Both Set A and B had same kind of questions parallel to each other with different values					
Typology	**No. of Items**	**Marks**	**Competency**	**Difficulty Level**	**Cognitive Level**	**Competencies Required**
MC	2	4	Number knowledge	Easy	Knowledge	Choose biggest number 2 digit & 3 digit
MC	1	2	Number knowledge	Medium	Knowledge	Choose smallest number 4 digit
MC	3	6	Fractions	Medium	Knowledge	Choose biggest & smallest fraction of same and different denominator
MC	2	4	Time	Easy	Knowledge	Reading time from a clock
MC	2	4	Geometry	Difficult	Knowledge	Triangle type (based on angles) - name
Answer	2	4	Fractions	Easy	Knowledge	Shaded figure write fraction
Answer	1	2	Fractions	Medium	Comprehension	Numerical addition with same denominator
Answer	1	2	Fractions	Difficult	Comprehension	Numerical addition with diff denominator
Answer	1	2	Geometry	Difficult	Knowledge	Write angle
Answer	1	2	Number line	Medium	Knowledge	Integers (negative & positive numbers)
Answer	1	2	Time	Medium	Comprehension	Calculation of time
Answer	3	6	Operations	Easy	Comprehension	Numerical 3 digit addition with carry; 4 digit subtraction with no borrow; 3 by 2 digit multiplication
Answer	1	2	Operations	Difficult	Comprehension	Numerical decimal subtraction with borrow
Answer	2	4	Operations	Medium	Comprehension	Numerical 3 by 1 digit division no remainder; 3 by 1 digit division remainder
Answer	1	2	Operations	Difficult	Comprehension	Numerical expansion using BODMAS
Answer	3	6	Operations	Easy	Application	Word problem 3 and 3 digit addition; 2 digit subtraction with borrow; 2 by 2 digit multiplication
Answer	1	2	Operations	Difficult	Application	Word problem 3 by 1 digit division
Answer	1	2	Area	Difficult	Application	Word problem 2 by 1
Answer	1	2	Perimeter	Medium	Application	Word problem
Answer	1	2	Perimeter & cost	Difficult	Application	Word problem
Answer	2	4	Application	Medium	Application	Word problem - calendar
Answer	2	4	Application	Medium	Application	Word problem - menu
Answer	2	4	Graph-pictograph	Medium	Application	Interpretation of graph
Answer	4	8	Table interpretation	Difficult	Application	Interpretation of table

Convergent validity

It refers to the degree to which a test is correlated with other test that it is theoretically predicted to correlate with. Parallel form correlation for reliability achieved a coefficient value of 0.97 which is a very high value indicating high reliability of the test forms A and B. It indicates that as many times both these are administered made under the same frame defining the construct (math) on the students of same age, Std and competency the results will be 97% similar, until and unless some intervention is given to improve the learning levels. It is also indicative of the fact that the items in the forms A and B are highly similar. They are made out of one frame and are generating same results. The correlation between two forms made out of the same frame also is a measure of validity called convergent validity. This depicts that the items in both the assessments have been made out of the same and pre defined frame which defines the construct. The tests are similar and convergent to each other and they are measuring what they intend to measure.

Criterion validity

It involves the correlation between the test and a criterion variable (or variables) taken as representative of the construct. In other words, it compares the test with other measures or outcomes (the criteria) already held to be valid. If both are giving same results either in a good or bad way then it shows validity by being able to match up to the criterion. Criterion validity is further of two kinds – concurrent and predictive. If two assessments measuring same things are administered at the same time to people in same age group and with similar competency then there is a high probability of high positive correlation. This situation is concurrent validity. In Bihar math assessment there is a high correlation between both the forms A and B as seen in the earlier section of the report. This means concurrently two assessments are indicating same kind of results as they have been made out of the same frame for same purpose. Predictive validity is slightly different because some other criterion of the construct which is predictive of the latter is pre - utilized to gauge what might happen later on when the original assessment will be applied. Here those students who have performed poorly on the reading assessment or are at the beginner level have also performed poorly on items which require reading, understanding and applying the knowledge to form or pick an appropriate answer. Reading is a prerequisite not only for a language assessment but for any subject assessment as the student can respond correctly only when he himself can read the items correctly and make sense out of it.

Table 3.4 Qualitative analysis for criterion validity- Bihar mathematics assessment Std 6

		Assessment Details		All children tested	% Children answering questions correctly					Prerequisite
		Std 6 Math : Pen-paper test - May 2014			Std 2 level readers (story)	Std 1 level readers (Para)	Word level readers	Letter level readers	Beginner level readers	
Number knowledge	Q1	2 digit	Comparison	79.4	89.9	74.6	61.1	46.8	36.7	Reading + Number Knowledge
Number knowledge	Q2	3 digit	Comparison	74.5	85.3	69.4	56.1	40.2	33.4	
Number knowledge	Q3	4digit	Comparison	69.9	81.8	62.5	49	35.5	30.2	
Operations	Q13a	3 digit	Addition with carrying	81.9	88.8	80.8	70.7	59.2	42.7	Reading + Operations
Operations	Q13b	4digit	Subtraction without borrowing	78	84.8	75.6	67.7	57	39.9	
Operations	Q13c	decimals	Subtraction with borrowing	60.2	69.4	54.6	44.2	34.2	24	
Operations	Q13d	3digit by 2 digit	Multiplication	59.5	69.7	53.5	41.7	29.4	22	
Operations	Q13e	3 digit by 1 digit	Division	40.5	49.7	32.9	24.1	16.4	14.2	
Operations	Q13f	3 digit by 1 digit	Division	54.2	65.4	47.4	33.6	21.6	17	
Operations	Q13g	1 digit	BODMAS	26.3	32.8	20.1	14.8	9.6	10.5	
Operations	Q14	3 digit	Word problem addition	64.7	76.7	57.1	44	31.6	25.9	
Operations	Q15	3 digit	Word problem subtraction	53.6	65.7	44	28.8	24.8	18.6	
Operations	Q16	2 digit	Word problem multiplication	34.4	43.5	24.9	18.5	13.7	19.8	
Operations	Q17	3 digit by 1 digit	Word problem division	23.5	30.4	15.8	10.5	8.7	6.8	
Fractions	Q4	Same denominator	Comparison	71.5	81	67.2	52.8	43.1	35.5	Reading + Fractions
Fractions	Q5	Different denominator	Comparison	12.5	14.1	11	9.8	8.6	9.3	
Fractions	Q6	Different denominator		14.9	14.9	15.8	16.6	13	8.8	
Fractions	Q13h	Same denominator	Addition	49.1	60.3	39.9	29.1	19.1	16.8	
Fractions	Q13i	Different denominator	Subtraction	37.4	47.9	27.8	18.1	11.9	9.8	
Integer	Q9	Number line	Postitive & negative numbers	18.9	24.4	12.6	9.8	6.8	5	Reading + Integers
Geometry	Q8a	Triangle	Characteristics	36.7	42.2	31.6	27.4	23.1	21.3	Reading + Geometry
Geometry	Q8b	Triangle	Characteristics	30	34.7	25.6	23.3	17.8	15.3	
Geometry	Q12	Angles	Recognizing	25	30.5	20.4	14.6	11.1	10	
Application	Q7	Time calculation	Word problem with time	46	54	37.5	33.1	47.2	22.3	Reading + Application
Application	Q18a	Area problem	Visual image	29.9	36.8	23.7	17.2	12.5	11.7	
Application	Q18b	Perimeter	Visual image	21.6	27.2	15.3	13.7	7.7	8.4	
Application	Q18c	Calculation	Word problem	13.4	16.8	9.8	7.9	5.2	6.1	
Application	Q11a	Clock	Telling time	22.3	26.9	18.7	14.4	9.4	9.6	
Application	Q11b	Clock	Telling time	45.5	55	38.5	28.5	21.3	15.4	
Application	Q19a	Calendar	Word problem 1	30.5	39.1	21.9	15.5	10.4	9.1	
Application	Q19b	Calendar	Word problem 2	22.9	29.5	15.6	12.3	6.7	5.4	
Application	Q20a	Menu card	Menu word problem 1	34.8	44.1	24.2	18.8	13.3	12.6	
Application	Q20b	Menu card	Menu word problem 2	14.3	18.3	9.7	8.1	5.5	4.2	
Application	Q21a	Pictograph	Pictograph word problem 1	55.4	69	42.6	30	22.6	19.9	
Application	Q21b	Pictograph	Pictograph word problem 2	15.4	19.4	11	10.4	5.7	4.2	
Application	Q22a	Data table	Data table word problem 1	42.2	54.3	29.3	20.2	15.4	11.9	
Application	Q22b	Data table	Data table word problem 2	38	49.1	25.5	18.9	13.4	12.9	
Application	Q22c	Data table	Data table word problem 3	19.4	25	13.7	9.1	6.7	5.6	
Application	Q22d	Data table	Data table word problem 4	16.8	22.2	10.5	7.6	5.1	5.4	
		Total children tested		17640						

Inter item correlation[7]

Inter item correlation shows how well do the items internally correlate and how consistent they are in predicting each other's performance. The items which have a negative correlation with most of the other items show that there is some problem of internal consistency. The possible reasons could be that the content is different or construct has too many contexts and contents to be assessed or there is language inconsistency or difficulty level inconsistency or items are assessing completely different things. A negative correlation is not good for any item with other item as it means performance on one item would be different from another item and hence they are not predictive. A zero correlation is still better as it means there is no correlation between and effect on performance on the item. Positive correlations are ideal. More the negative correlations the worse the item becomes in internal consistency issues and predictability. This is a measure of internal consistency of the test. Internal consistency also means reliability as well as validity. In the study we find that none of the items negatively correlate with each other. On finding the **Average inter item** correlation the value indicates 0.22 which is in positive direction and a good indication of none the items being completely different from the whole test or the frame of the construct defined.

Item total correlation[8]

This correlation is the regressive correlation done between the scoring data of each of the items situated in different columns and the total raw score in the last column for all the students on all the items. In this concept a higher correlation means that performance of all the students on one item is predictive of their performance on remaining items or the test. A zero correlation means performance on that item does not predict or impact performance on the remaining items or the test. This is again a measure of internal consistency in the test. Internal consistency also means reliability as well as validity. Here we see that all items are positively correlated with the total and some items have been highlighted because the correlation was less than .50 but even on doing further analysis it was found that removing any item would reduce the quality of the test by lowering the consistency. Hence all these items together make a valid assessment.

[7]Inter item correlation shows both reliability and validity. It is a measure of internal consistency. It shows that scores in one item are matching and predicting with the remaining items of the assessment as well as whether all items are in sync with each other and measuring what they intend to do as a whole within the defined construct.
[8]Item total correlation is a measure of both reliability and validity. It shows whether there is a similar trend and positive correlation between responses of one item with the total result of the individual or the group. Any item would be predictor for the rest of the items thus. It also shows whether all items are performing in sync with the rest of the items as a whole and falls within the defined limit of the measurable construct or not.

Table 3.5 Inter item correlation- Bihar mathematics assessment Std 6																				
Items	q1	q2	q3	q4	q5	q6	q7	q8a	q8b	q9	q10a	q10b	q11a	q11b	q12	q13a	q13b	q13c	q13d	q13e	q13f
q1	1.00	0.66	0.57	0.46	0.09	0.09	0.31	0.20	0.19	0.18	0.22	0.22	0.16	0.26	0.18	0.25	0.24	0.25	0.25	0.22	0.27
q2	0.66	1.00	0.53	0.44	0.10	0.08	0.29	0.19	0.19	0.17	0.22	0.21	0.17	0.26	0.18	0.24	0.23	0.24	0.26	0.23	0.27
q3	0.57	0.53	1.00	0.41	0.11	0.01	0.25	0.21	0.18	0.17	0.22	0.21	0.15	0.25	0.18	0.22	0.22	0.23	0.26	0.23	0.27
q4	0.46	0.44	0.41	1.00	0.01	0.05	0.23	0.19	0.16	0.14	0.18	0.18	0.13	0.21	0.13	0.20	0.20	0.20	0.22	0.19	0.23
q5	0.09	0.10	0.11	0.01	1.00	0.35	0.16	0.17	0.21	0.15	0.14	0.14	0.15	0.13	0.22	0.05	0.05	0.09	0.06	0.12	0.07
q6	0.09	0.08	0.01	0.05	0.35	1.00	0.15	0.11	0.14	0.07	0.10	0.08	0.08	0.07	0.13	0.02	0.02	0.03	0.02	0.05	0.02
q7	0.31	0.29	0.25	0.23	0.16	0.15	1.00	0.21	0.22	0.18	0.20	0.19	0.18	0.23	0.21	0.14	0.14	0.18	0.17	0.19	0.19
q8a	0.20	0.19	0.21	0.19	0.17	0.11	0.21	1.00	0.28	0.17	0.17	0.16	0.17	0.16	0.25	0.10	0.09	0.11	0.11	0.12	0.12
q8b	0.19	0.19	0.18	0.16	0.21	0.14	0.22	0.28	1.00	0.19	0.18	0.18	0.18	0.17	0.27	0.09	0.09	0.12	0.10	0.14	0.12
q9	0.18	0.17	0.17	0.14	0.15	0.07	0.18	0.17	0.19	1.00	0.26	0.26	0.20	0.21	0.26	0.11	0.12	0.16	0.16	0.20	0.17
q10a	0.22	0.22	0.22	0.18	0.14	0.10	0.20	0.17	0.18	0.26	1.00	0.72	0.22	0.25	0.25	0.15	0.15	0.19	0.19	0.20	0.19
q10b	0.22	0.21	0.21	0.18	0.14	0.08	0.19	0.16	0.18	0.26	0.72	1.00	0.22	0.25	0.25	0.14	0.15	0.19	0.18	0.21	0.19
q11a	0.16	0.17	0.15	0.13	0.15	0.08	0.18	0.17	0.18	0.20	0.22	0.22	1.00	0.32	0.26	0.11	0.11	0.14	0.12	0.15	0.15
q11b	0.26	0.26	0.25	0.21	0.13	0.07	0.23	0.16	0.17	0.21	0.25	0.25	0.32	1.00	0.24	0.19	0.18	0.21	0.21	0.21	0.24
q12	0.18	0.18	0.18	0.13	0.22	0.13	0.21	0.25	0.27	0.26	0.25	0.25	0.26	0.24	1.00	0.11	0.13	0.16	0.13	0.18	0.16
q13a	0.25	0.24	0.22	0.20	0.05	0.02	0.14	0.10	0.09	0.11	0.15	0.14	0.11	0.19	0.11	1.00	0.37	0.31	0.30	0.23	0.28
q13b	0.24	0.23	0.22	0.20	0.05	0.02	0.14	0.09	0.09	0.12	0.15	0.15	0.11	0.18	0.13	0.37	1.00	0.37	0.30	0.23	0.27
q13c	0.25	0.24	0.23	0.20	0.09	0.03	0.18	0.11	0.12	0.16	0.19	0.19	0.14	0.21	0.16	0.31	0.37	1.00	0.32	0.28	0.30
q13d	0.25	0.26	0.26	0.22	0.06	0.02	0.17	0.11	0.10	0.16	0.19	0.18	0.12	0.21	0.13	0.30	0.30	0.32	1.00	0.31	0.37
q13e	0.22	0.23	0.23	0.19	0.12	0.05	0.19	0.12	0.14	0.20	0.20	0.21	0.15	0.21	0.18	0.23	0.23	0.28	0.31	1.00	0.39
q13f	0.27	0.27	0.27	0.23	0.07	0.02	0.19	0.12	0.12	0.17	0.19	0.19	0.15	0.24	0.16	0.28	0.27	0.30	0.37	0.39	1.00
q13g	0.19	0.19	0.20	0.14	0.17	0.08	0.20	0.18	0.21	0.26	0.24	0.24	0.21	0.23	0.30	0.15	0.15	0.21	0.20	0.24	0.23
q13h	0.28	0.28	0.28	0.24	0.12	0.05	0.23	0.16	0.18	0.24	0.26	0.27	0.19	0.29	0.24	0.26	0.23	0.29	0.31	0.32	0.39
q13i	0.25	0.26	0.26	0.22	0.13	0.05	0.23	0.15	0.18	0.25	0.26	0.26	0.19	0.27	0.24	0.22	0.21	0.27	0.29	0.33	0.35
q14	0.33	0.32	0.32	0.27	0.08	0.03	0.23	0.16	0.14	0.19	0.22	0.22	0.18	0.28	0.19	0.27	0.25	0.27	0.30	0.27	0.32
q15	0.30	0.30	0.30	0.24	0.10	0.03	0.26	0.16	0.16	0.20	0.23	0.22	0.19	0.29	0.20	0.24	0.23	0.29	0.30	0.29	0.33
q16	0.23	0.23	0.25	0.18	0.15	0.06	0.25	0.17	0.19	0.25	0.23	0.23	0.22	0.28	0.25	0.18	0.18	0.23	0.25	0.26	0.27
q17	0.19	0.20	0.21	0.17	0.17	0.09	0.24	0.19	0.21	0.25	0.23	0.23	0.22	0.26	0.29	0.15	0.15	0.20	0.20	0.24	0.23
q18a	0.21	0.20	0.21	0.16	0.18	0.10	0.23	0.22	0.24	0.27	0.26	0.25	0.23	0.25	0.34	0.16	0.16	0.20	0.20	0.22	0.22
q18b	0.18	0.18	0.18	0.13	0.21	0.13	0.22	0.21	0.26	0.26	0.24	0.25	0.22	0.23	0.34	0.13	0.14	0.17	0.16	0.21	0.19
q18c	0.13	0.13	0.14	0.10	0.20	0.12	0.19	0.20	0.22	0.20	0.19	0.19	0.21	0.19	0.29	0.10	0.10	0.13	0.12	0.15	0.13
q19a	0.22	0.22	0.23	0.17	0.17	0.09	0.25	0.19	0.20	0.25	0.24	0.24	0.21	0.26	0.29	0.16	0.15	0.20	0.20	0.23	0.23
q19b	0.18	0.19	0.19	0.15	0.14	0.07	0.21	0.18	0.20	0.23	0.22	0.22	0.22	0.24	0.28	0.14	0.13	0.17	0.17	0.22	0.20
q20a	0.23	0.23	0.24	0.19	0.14	0.07	0.25	0.17	0.18	0.23	0.23	0.22	0.22	0.27	0.24	0.17	0.17	0.21	0.20	0.23	0.24
q20b	0.14	0.14	0.15	0.11	0.16	0.11	0.19	0.13	0.16	0.19	0.18	0.18	0.18	0.20	0.21	0.11	0.11	0.15	0.12	0.19	0.15
q21a	0.32	0.32	0.31	0.27	0.10	0.02	0.24	0.18	0.19	0.23	0.25	0.25	0.19	0.30	0.22	0.23	0.22	0.24	0.26	0.25	0.30
q21b	0.13	0.14	0.14	0.11	0.17	0.11	0.16	0.14	0.19	0.20	0.19	0.18	0.16	0.17	0.21	0.11	0.10	0.14	0.12	0.16	0.14
q22a	0.27	0.28	0.28	0.22	0.12	0.04	0.24	0.18	0.18	0.25	0.25	0.24	0.20	0.29	0.23	0.20	0.19	0.23	0.24	0.25	0.29
q22b	0.25	0.26	0.27	0.22	0.13	0.05	0.25	0.18	0.20	0.25	0.26	0.25	0.20	0.28	0.26	0.19	0.19	0.23	0.23	0.25	0.26
q22c	0.15	0.15	0.14	0.14	0.12	0.04	0.15	0.11	0.14	0.16	0.14	0.17	0.17	0.19	0.19	0.13	0.11	0.12	0.12	0.16	0.15
q22d	0.16	0.17	0.18	0.14	0.20	0.12	0.21	0.19	0.24	0.26	0.23	0.23	0.22	0.24	0.30	0.12	0.12	0.17	0.15	0.21	0.18
Total score	0.53	0.53	0.52	0.44	0.31	0.19	0.47	0.38	0.41	0.45	0.50	0.49	0.41	0.51	0.49	0.40	0.39	0.46	0.46	0.49	0.51

Items	q13g	q13h	q13i	q14	q15	q16	q17	q18a	q18b	q18c	q19a	q19b	q20a	q20b	q21a	q21b	q22a	q22b	q22c	q22d	Total score
q1	0.19	0.28	0.25	0.33	0.30	0.23	0.19	0.21	0.18	0.13	0.22	0.18	0.23	0.14	0.32	0.13	0.27	0.25	0.15	0.16	0.53
q2	0.19	0.28	0.26	0.32	0.30	0.23	0.20	0.20	0.18	0.13	0.22	0.19	0.23	0.14	0.32	0.14	0.28	0.26	0.15	0.17	0.53
q3	0.20	0.28	0.26	0.32	0.30	0.25	0.21	0.21	0.18	0.14	0.23	0.19	0.24	0.15	0.31	0.14	0.28	0.27	0.14	0.18	0.52
q4	0.14	0.24	0.22	0.27	0.24	0.18	0.17	0.16	0.13	0.10	0.17	0.15	0.19	0.11	0.27	0.11	0.22	0.22	0.14	0.14	0.44
q5	0.17	0.12	0.13	0.08	0.10	0.15	0.17	0.18	0.21	0.20	0.17	0.14	0.14	0.16	0.10	0.17	0.12	0.13	0.12	0.20	0.31
q6	0.08	0.05	0.05	0.03	0.03	0.06	0.09	0.10	0.13	0.12	0.09	0.07	0.07	0.11	0.02	0.11	0.04	0.05	0.04	0.12	0.19
q7	0.20	0.23	0.23	0.23	0.26	0.25	0.24	0.23	0.22	0.19	0.25	0.21	0.25	0.19	0.24	0.16	0.24	0.25	0.15	0.21	0.47
q8a	0.18	0.16	0.15	0.16	0.16	0.17	0.19	0.22	0.21	0.20	0.19	0.18	0.17	0.13	0.18	0.14	0.18	0.18	0.11	0.19	0.38
q8b	0.21	0.18	0.18	0.14	0.16	0.19	0.21	0.24	0.26	0.22	0.20	0.20	0.18	0.16	0.19	0.19	0.18	0.20	0.14	0.24	0.41
q9	0.26	0.24	0.25	0.19	0.20	0.25	0.25	0.27	0.26	0.20	0.25	0.23	0.23	0.19	0.23	0.20	0.25	0.25	0.16	0.26	0.45
q10a	0.24	0.26	0.26	0.22	0.23	0.23	0.23	0.26	0.24	0.19	0.24	0.22	0.23	0.18	0.25	0.19	0.25	0.26	0.14	0.23	0.50
q10b	0.24	0.27	0.26	0.22	0.22	0.23	0.23	0.25	0.25	0.19	0.24	0.22	0.22	0.18	0.25	0.18	0.24	0.25	0.17	0.23	0.49
q11a	0.21	0.19	0.19	0.18	0.19	0.22	0.22	0.23	0.22	0.21	0.21	0.22	0.22	0.18	0.19	0.16	0.20	0.20	0.17	0.22	0.41
q11b	0.23	0.29	0.27	0.28	0.29	0.28	0.26	0.25	0.23	0.19	0.26	0.24	0.27	0.20	0.30	0.17	0.29	0.28	0.19	0.24	0.51
q12	0.30	0.24	0.24	0.19	0.20	0.25	0.29	0.34	0.34	0.29	0.29	0.28	0.24	0.21	0.22	0.21	0.23	0.26	0.19	0.30	0.49
q13a	0.15	0.26	0.22	0.27	0.24	0.18	0.15	0.16	0.13	0.10	0.16	0.14	0.17	0.11	0.23	0.11	0.20	0.19	0.13	0.12	0.40
q13b	0.15	0.23	0.21	0.25	0.23	0.18	0.15	0.16	0.14	0.10	0.15	0.13	0.17	0.11	0.22	0.10	0.19	0.19	0.11	0.12	0.39
q13c	0.21	0.29	0.27	0.27	0.29	0.23	0.20	0.20	0.17	0.13	0.20	0.17	0.21	0.15	0.24	0.14	0.23	0.23	0.12	0.17	0.46
q13d	0.20	0.31	0.29	0.30	0.30	0.25	0.20	0.20	0.16	0.12	0.20	0.17	0.20	0.12	0.26	0.12	0.24	0.23	0.12	0.15	0.46
q13e	0.24	0.32	0.33	0.27	0.29	0.26	0.24	0.22	0.21	0.15	0.23	0.22	0.23	0.19	0.25	0.16	0.25	0.25	0.16	0.21	0.49
q13f	0.23	0.39	0.35	0.32	0.33	0.27	0.23	0.22	0.19	0.13	0.23	0.20	0.24	0.15	0.30	0.14	0.29	0.26	0.15	0.18	0.51
q13g	1.00	0.38	0.36	0.23	0.25	0.27	0.29	0.33	0.32	0.23	0.27	0.26	0.27	0.21	0.25	0.20	0.26	0.27	0.19	0.29	0.51
q13h	0.38	1.00	0.57	0.34	0.35	0.32	0.31	0.30	0.26	0.18	0.28	0.25	0.29	0.20	0.34	0.20	0.32	0.31	0.20	0.24	0.59
q13i	0.36	0.57	1.00	0.32	0.34	0.33	0.30	0.32	0.28	0.20	0.31	0.27	0.29	0.21	0.33	0.20	0.32	0.32	0.21	0.26	0.59
q14	0.23	0.34	0.32	1.00	0.52	0.37	0.30	0.27	0.23	0.18	0.29	0.23	0.31	0.19	0.37	0.17	0.33	0.33	0.17	0.20	0.56
q15	0.25	0.35	0.34	0.52	1.00	0.44	0.35	0.27	0.25	0.19	0.30	0.25	0.33	0.21	0.38	0.19	0.35	0.36	0.20	0.24	0.59
q16	0.27	0.32	0.33	0.37	0.44	1.00	0.49	0.33	0.30	0.24	0.34	0.31	0.36	0.28	0.34	0.23	0.34	0.36	0.24	0.31	0.59
q17	0.29	0.31	0.30	0.30	0.35	0.49	1.00	0.33	0.32	0.26	0.34	0.33	0.33	0.29	0.29	0.24	0.31	0.33	0.22	0.34	0.56
q18a	0.33	0.30	0.32	0.27	0.27	0.33	0.33	1.00	0.60	0.38	0.37	0.33	0.31	0.24	0.29	0.24	0.30	0.32	0.21	0.33	0.57
q18b	0.32	0.26	0.28	0.23	0.25	0.30	0.32	0.60	1.00	0.52	0.37	0.33	0.30	0.25	0.25	0.25	0.27	0.30	0.21	0.35	0.55
q18c	0.23	0.18	0.20	0.18	0.19	0.24	0.26	0.38	0.52	1.00	0.31	0.28	0.25	0.25	0.19	0.24	0.20	0.23	0.18	0.29	0.44
q19a	0.27	0.28	0.31	0.29	0.30	0.34	0.34	0.37	0.37	0.31	1.00	0.48	0.38	0.29	0.34	0.26	0.35	0.37	0.20	0.33	0.57
q19b	0.26	0.25	0.27	0.23	0.25	0.31	0.33	0.33	0.33	0.28	0.48	1.00	0.33	0.30	0.30	0.27	0.32	0.33	0.24	0.35	0.53
q20a	0.27	0.29	0.29	0.31	0.33	0.36	0.33	0.31	0.30	0.25	0.38	0.33	1.00	0.41	0.41	0.26	0.39	0.41	0.27	0.32	0.57
q20b	0.21	0.20	0.21	0.19	0.21	0.28	0.29	0.24	0.25	0.25	0.29	0.30	0.41	1.00	0.24	0.31	0.26	0.28	0.21	0.31	0.44
q21a	0.25	0.34	0.33	0.37	0.38	0.34	0.29	0.29	0.25	0.19	0.34	0.30	0.41	0.24	1.00	0.26	0.45	0.42	0.28	0.28	0.60
q21b	0.20	0.20	0.20	0.17	0.19	0.23	0.24	0.24	0.25	0.24	0.26	0.27	0.26	0.31	0.26	1.00	0.23	0.26	0.21	0.28	0.42
q22a	0.26	0.32	0.32	0.33	0.35	0.34	0.31	0.30	0.27	0.20	0.35	0.32	0.39	0.26	0.45	0.23	1.00	0.54	0.33	0.33	0.60
q22b	0.27	0.31	0.32	0.33	0.36	0.36	0.33	0.32	0.30	0.23	0.37	0.33	0.41	0.28	0.42	0.26	0.54	1.00	0.33	0.38	0.60
q22c	0.19	0.20	0.21	0.17	0.20	0.24	0.22	0.21	0.21	0.18	0.20	0.24	0.27	0.21	0.28	0.21	0.33	0.33	1.00	0.33	0.41
q22d	0.29	0.24	0.26	0.20	0.24	0.31	0.34	0.33	0.35	0.29	0.33	0.35	0.32	0.31	0.28	0.28	0.33	0.38	0.33	1.00	0.52
Total score	0.51	0.59	0.59	0.56	0.59	0.59	0.56	0.57	0.55	0.44	0.57	0.53	0.57	0.44	0.60	0.42	0.60	0.60	0.41	0.52	1.00

Table 3.6 Item total correlation- Bihar mathematics assessment Std 6		
Item	Item total correlation	Corrected item total correlation
q1	0.52	0.49
q2	0.52	0.49
q3	0.51	0.48
q4	0.43	0.39
q5	0.33	0.28
q6	0.21	0.16
q7	0.47	0.43
q8a	0.38	0.34
q8b	0.41	0.37
q9	0.46	0.42
q10a	0.49	0.45
q10b	0.49	0.45
q11a	0.42	0.37
q11b	0.51	0.47
q12	0.49	0.46
q13a	0.40	0.36
q13b	0.39	0.35
q13c	0.46	0.42
q13d	0.46	0.41
q13e	0.48	0.44
q13f	0.5	0.46
q13g	0.51	0.47
q13h	0.58	0.55
q13i	0.58	0.54
q14	0.56	0.52
q15	0.58	0.54
q16	0.58	0.55
q17	0.56	0.52
q18a	0.57	0.54
q18b	0.56	0.52
q18c	0.46	0.42
q19a	0.58	0.54
q19b	0.53	0.49
q20a	0.57	0.54
q20b	0.46	0.42
q21a	0.59	0.56
q21b	0.43	0.39
q22a	0.59	0.56
q22b	0.6	0.57
q22c	0.42	0.37
q22d	0.54	0.5

Item level analysis

This includes item information, item person map, item difficulty, Distracter analysis of the MCI, Item discrimination, Item characteristic curves for the items with strange performance

Item difficulty:

The term item difficulty or p value expresses the proportion or percentage of students who answered the item correctly. Item difficulty can range from 0.0 (none of the students answered the item correctly) to 1.0 (all of the students answered the item correctly). Item difficulty is also termed as Item facility.

Figure 3.1 Understanding item difficulty										
0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Nobody is able to solve this item correctly		20%	30%	40%	50 percent could solve this item	60%	70%	80%	Everybody is able to solve this item correctly	
Very Difficult item		Difficult	Difficult	Medium	Medium	Medium	Easy	Easy	Very easy item	

The normal distribution curve has a spread of population of all kinds i.e. less able to average to more able. As per the Rasch model of item analysis if the item difficulty and the person ability are calibrated on the same scale it would appear as the figure given below.

Figure 3.2 Understanding item difficulty and person ability																					
Location of Less able group										**Average ability group**	Location of More able group										
-1.0	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	+0.1	+0.2	+0.3	+0.4	+0.5	+0.6	+0.7	+0.8	+0.9	+1.0	
Negative sided S.D. or standard deviation from mean										MEAN	Positive sided S.D. or standard deviation from mean										
Location of Easier items										Medium items	Location of Difficult items										

Table 3.7 Item difficulty set wise- Bihar mathematics assessment Std 6

Set A	p value 1	Set B	p value 2	Difficulty Level
q1_1	0.79	q1_2	0.8	Both Easy
q2_1	0.75	q2_2	0.74	Both Easy
q3_1	0.72	q3_2	0.68	Set A Easy and B Medium
q4_1	0.71	q4_2	0.73	Both Easy
q5_1	0.13	q5_2	0.12	Both Difficult
q6_1	0.2	q6_2	0.1	Both Difficult
q7_1	0.47	q7_2	0.46	Both Medium
q8a_1	0.39	q8a_2	0.34	Both Difficult
q8b_1	0.3	q8b_2	0.3	Both Difficult
q9_1	0.19	q9_2	0.19	Both Difficult
q10a_1	0.42	q10a_2	0.34	Set A is Medium and B is Difficult
q10b_1	0.35	q10b_2	0.35	Both Difficult
q11a_1	0.21	q11a_2	0.24	Both Difficult
q11b_1	0.46	q11b_2	0.46	Both Medium
q12_1	0.25	q12_2	0.25	Both Difficult
q13a_1	0.82	q13a_2	0.82	Both Easy
q13b_1	0.79	q13b_2	0.78	Both Easy
q13c_1	0.61	q13c_2	0.59	Both Medium
q13d_1	0.62	q13d_2	0.57	Both Medium
q13e_1	0.41	q13e_2	0.4	Both Medium
q13f_1	0.53	q13f_2	0.55	Both Medium
q13g_1	0.27	q13g_2	0.26	Both Difficult
q13h_1	0.49	q13h_2	0.5	Both Medium
q13i_1	0.38	q13i_2	0.38	Both Difficult
q14_1	0.66	q14_2	0.64	Both Medium
q15_1	0.54	q15_2	0.54	Both Medium
q16_1	0.34	q16_2	0.36	Both Difficult
q17_1	0.23	q17_2	0.24	Both Difficult
q18a_1	0.3	q18a_2	0.3	Both Difficult
q18b_1	0.21	q18b_2	0.22	Both Difficult
q18c_1	0.14	q18c_2	0.13	Both Difficult
q19a_1	0.34	q19a_2	0.27	Both Difficult
q19b_1	0.22	q19b_2	0.24	Both Difficult
q20a_1	0.34	q20a_2	0.36	Both Difficult
q20b_1	0.15	q20b_2	0.14	Both Difficult
q21a_1	0.53	q21a_2	0.59	Both Medium
q21b_1	0.15	q21b_2	0.17	Both Difficult
q22a_1	0.44	q22a_2	0.4	Both Medium
q22b_1	0.39	q22b_2	0.38	Both Difficult
q22c_1	0.09	q22c_2	0.3	Both Difficult
q22d_1	0.17	q22d_2	0.17	Both Difficult
Total Difficult Items	23	Total Difficult Items	24	
Total Medium Items	12	Total Medium Items	12	
Total Easy Items	6	Total Easy Items	5	
Totals	41		41	

Index

p-values	Indicator	Difficulty Level
0.0 to 0.39		Difficult
0.4 to 0.69		Medium
0.7 to 1.0		Easy

Important

The items which have the biggest p value in Set A and B are marked in Yellow which denotes that these were attempted by large number of people and were comparatively easy.

The items which have the smallest p value in Set A and B are marked in Red which denotes that these were attempted by least number of people and were comparatively difficult.

Item discrimination

What is Findlay's Index and Item discrimination

We want items on our tests that inform us about student learning. We want items that differentiate or discriminate between those who score high on the test and those who score low on the test. We want this index to be easy to compute and easy to understand. The Findlay Index is the easiest to use. There are three kinds of items.

- Over Discriminating items – The poor students do worse and good students do more than expected.

- Under Discriminating items – The good students perform worse or less than expectation and the poor students surprisingly do better. These are invalid items.

- Discriminating item – Students perform according to their caliber and probability* to get the answer correct.

Findlay Index formula:

Rank order students in descending order based on total test score.

Split these students into three groups:

NRU = Number of students getting the item correct in the upper band of the test

NRL = Number of students getting the item correct in the lower band of the test

NU = Number of total students in the upper band of the test

(Ignore the middle band of students)

$$\text{Findlay Index} = \frac{\text{No. correct in Upper band} - \text{No. correct in Lower band}}{\text{No. of students in Upper band}}$$

The number in the upper and lower bands must be the same.

Figure 3.3 Understanding item discrimination			
DISCRIMINATION TYPE	OBSERVED RESPONSE	EXPECTED RESPONSE	ACCEPTANCE
OVER DISCRIMINATING	GOOD STUDENTS DID EVEN BETTER THAN EXPECTED AND POOR STUDENTS DID WORSE THAN EXPECTED	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	x
NORMAL DISCRIMINATING	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	✓
UNDER DISCRIMINATING	GOOD STUDENTS DID WORST THAN EXPECTED AND POOR STUDENTS DID MUCH BETTER THAN EXPECTED	THE PROBABILITY OF PERFORMING BETTER INCREASED WITH THE ABILITY AND VICE VERSA	x

Item	Difficulty	Interpretation	Discrimination	Discrimination
1	0.79	Easy	0.49	Students are not performing as per their ability
2	0.74	Easy	0.48	Students are not performing as per their ability
3	0.7	Easy	0.47	Students are not performing as per their ability
4	0.71	Easy	0.39	Students are not performing as per their ability
5	0.12	Most Difficult	0.27	Students are not performing as per their ability
6	0.14	Difficult	0.14	Students are not performing as per their ability
7	0.46	Medium	0.42	Students are not performing as per their ability
8a	0.36	Difficult	0.33	Students are not performing as per their ability
8b	0.3	Difficult	0.36	Students are not performing as per their ability
9	0.19	Difficult	0.41	Students are not performing as per their ability
10a	0.38	Difficult	0.45	Students are not performing as per their ability
10b	0.35	Difficult	0.45	Students are not performing as per their ability
11a	0.22	Difficult	0.37	Students are not performing as per their ability
11b	0.45	Medium	0.47	Students are not performing as per their ability
12	0.25	Difficult	0.44	Students are not performing as per their ability
13a	0.81	Most Easy	0.36	Students are not performing as per their ability
13b	0.78	Easy	0.35	Students are not performing as per their ability
13c	0.6	Medium	0.41	Students are not performing as per their ability
13d	0.59	Medium	0.42	Students are not performing as per their ability
13e	0.4	Medium	0.44	Students are not performing as per their ability
13f	0.54	Medium	0.46	Students are not performing as per their ability
13g	0.26	Difficult	0.47	Students are not performing as per their ability
13h	0.49	Medium	0.55	Students are performing as per their ability
13i	0.37	Difficult	0.54	Students are performing as per their ability
14	0.65	Medium	0.52	Students are performing as per their ability
15	0.53	Medium	0.54	Students are performing as per their ability
16	0.34	Difficult	0.55	Students are performing as per their ability
17	0.23	Difficult	0.52	Students are performing as per their ability
18a	0.29	Difficult	0.53	Students are performing as per their ability
18b	0.21	Difficult	0.51	Students are performing as per their ability
18c	0.13	Difficult	0.41	Students are not performing as per their ability
19a	0.3	Difficult	0.53	Students are performing as per their ability
19b	0.22	Difficult	0.49	Students are not performing as per their ability
20a	0.35	Difficult	0.53	Students are performing as per their ability
20b	0.14	Difficult	0.41	Students are not performing as per their ability
21a	0.55	Medium	0.56	Students are performing as per their ability
21b	0.15	Difficult	0.38	Students are not performing as per their ability
22a	0.42	Medium	0.55	Students are performing as per their ability
22b	0.38	Difficult	0.56	Students are performing as per their ability
22c	0.19	Difficult	0.37	Students are not performing as per their ability
22d	0.16	Difficult	0.49	Students are not performing as per their ability[9]

Table 3.8 A comprehensive table to understand the item function competency wise difficulty and discrimination- Bihar mathematics assessment Std 6

[9]Students are not performing as per their ability means either all students are over performing or underperforming. The item is not identifying whether the student is a top scorer or a bottom scorer. The difficult items have most of the time seen under performance by all students.

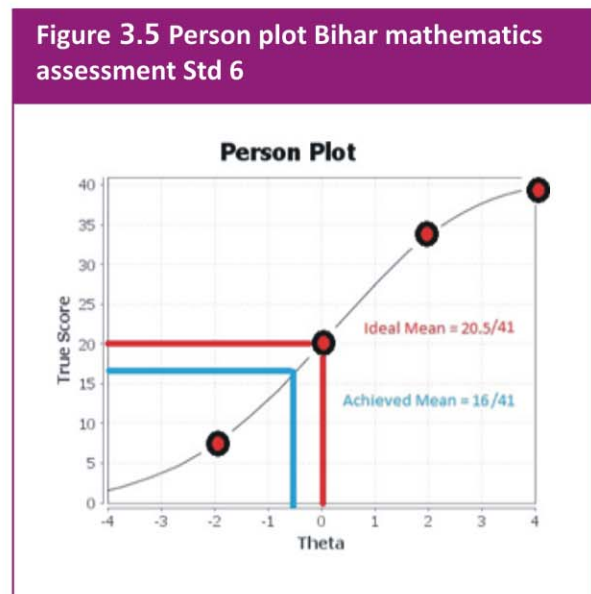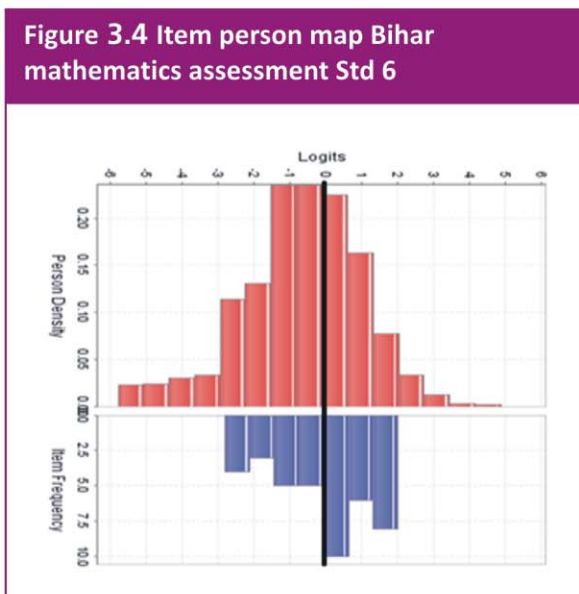| Table 3.9 Item location and corresponding performance- Bihar mathematics assessment Std 6			
Item	Location	Interpretation as per location on RASCH scale	Percentage of people attempting correctly
13a	-2.8	Easiest Item	81%
1	-2.58	Easy	79%
13b	-2.45	Easy	78%
2	-2.17	Easy	74%
4	-1.95	Easy	71%
3	-1.84	Easy	70%
14	-1.51	Easy	65%
13c	-1.21	Easy	60%
13d	-1.17	Easy	59%
21a	-0.94	Medium	55%
13f	-0.86	Medium	54%
15	-0.84	Medium	53%
13h	-0.58	Medium	49%
7	-0.4	Medium	46%
22a	-0.18	Medium	42%
13e	-0.08	Medium	40%
22b	0.05	Medium	38%
10a	0.06	Medium	38%
13i	0.09	Medium	37%
8a	0.14	Medium	36%
10b	0.25	Medium	35%
20a	0.25	Medium	35%
16	0.27	Medium	34%
11b*	0.38	Medium	45%
19a	0.52	Medium	30%
8b	0.55	Medium	30%
18a	0.56	Medium	29%
13g	0.8	Medium	26%
12	0.89	Medium	25%
17	1	Difficult	23%
19b	1.04	Difficult	22%
11a	1.09	Difficult	22%
18b	1.14	Difficult	21%
22c	1.32	Difficult	19%
9	1.36	Difficult	19%
22d	1.53	Difficult	16%
21b	1.67	Difficult	15%
6	1.73	Difficult	14%
20b	1.78	Difficult	14%
18c	1.88	Difficult	13%
5	1.97	Most Difficult	12%

Item Person Map

The given graph is known as the Item Person Map. It depicts the spread of item and persons on a common item-person scale. The mean location of item is 0.00 which means that average student here is located towards the negative side of the items where he or she can attempt the less difficult items with ease. The kurtosis for person map (peak of curve) is -0.66 and it is skewed (slope) at 0.22.

The distribution of the persons in the given figure reveal that more number of students are on the left side (less able than average) of the scale in comparison to the number of able students are towards the right side (more able than average). The difficulty levels of almost 14 - 15 of the items are matching with ability of the students close to the mean. But the 10 – 11 items which are all located below the mean are of very easy level for this group of people. Remaining 15 items are difficult even for the average student located closer to the mean.

There is a small group of people who are above the average group and away from the mean on the right side. All items are easy for them. There is a group who are located on the extreme negative side of the mean and found all the items difficult then their ability level. Remaining people are close to the mean on both sides and they found items of all categories as easy, medium to difficult.

Figure 3.4 Item person map Bihar mathematics assessment Std 6

Figure 3.5 Person plot Bihar mathematics assessment Std 6

Percentile bands and item benchmarking

If we see the student's performance with reference to what should have been their mean scores we find that very few students are in the highest percentile band. Where as if we see student's performance with reference to what has been their mean scores we find that there are almost equally distributed across all percentile bands. Also the last percentile band has a wide range of scores in comparison to the expected scores in the last percentile band. The shift is more towards the negative side of the mean in the real observed case. The mean in each percentile band and the mean within each one of these bands reflect the same. This overall reflects that the students have performed poorly in comparison to what was expected.

Table 3.10 Percentile bands for Bihar math assessment Std 6					
PERCENTAGE OF STUDENTS	25.30%	27.20%	24.40%	23.10%	Notice the increase in Percentage of students in the most able group when there is a broad range of scores to be achieved
PERCENTILE BAND	25TH	50TH	75TH	100TH	
RANGE OF PERCENTAGE OF MARKS	UPTO 25%	UPTO 50%	UPTO 75%	UPTO 100%	
MEAN SCORE WITHIN	0 to 9	10 TO 16	17 TO 23	23 TO 41	
ACTUAL MEAN	5.1	13.1	19.8	28.9	
ACTUAL TEST MEAN SCORE		16			Notice the shifted mean in Actual test mean score. Students on an average should have got the expected mean scores.
TEST SCORES	0 to 41				
EXPECTED TEST MEAN SCORE			20.5		
EXPECTED MEAN	5.7	15.4	24.9	33.8	
MEAN SCORE WITHIN	0 TO 10.25	10.25 TO 20.25	20.25 TO 30.75	30.75 TO 41	
RANGE OF PERCENTAGE OF MARKS	UPTO 25%	UPTO 50%	UPTO 75%	UPTO 100%	
PERCENTILE BAND	25TH	50TH	75TH	100TH	Notice the small Percentage of most able students who have performed as per expectation even when there is a narrow range of good scores to be achieved
PERCENTAGE OF STUDENTS	28.90%	38.60%	25.40%	7.10%	

Item benchmarking

All the students are put in specific percentile bands, then each of the items is seen in terms of the number of correct responses in each band. Those items in a particular percentile band which have percentage of students greater than or equal to 50% are suitable for such ability groups. Here we find that 8 items can't be bench marked which means 50% of the most able students also could not attempt these items correctly, 14 items are benchmarked at the most able level which means 50% of the most able students could attempt these items correctly, 10 items are benchmarked at the above average level, 7 items are benchmarked at the average level and 2 of the items are benchmarked at the low ability level. This shows that the students located on the extreme negative end of the mean and those who are in the 25th percentile band did not find many item suitable for them in terms of getting it attempted correctly.

Ability groups	Low Ability	Average	Above Average	Most Able	Benchmarked at which ability group
Items	25th Percentile	50th Percentile	75th Percentile	100th Percentile	
q1					
0	60	13.7	5	2.2	
1	40	86.3	95	97.8	Average
q2	%	%	%	%	
0	67	19.4	9.2	4.4	
1	33	80.6	90.8	95.6	Average
q3	%	%	%	%	
0	71	26.3	14.2	6.4	
1	29	73.7	85.8	93.6	Average
q4	%	%	%	%	
0	65	22.2	14.5	10.7	
1	35	77.8	85.5	89.3	Average
q5	%	%	%	%	
0	95.2	92.7	89.3	70.9	
1	4.8	7.3	10.7	29.1	can't be benchmarked
q6	%	%	%	%	
0	90.4	87.3	87.1	74.7	
1	9.6	12.7	12.9	25.3	can't be benchmarked
q7	%	%	%	%	
0	83	63.8	43.4	21.8	
1	17	36.2	56.6	78.2	Above Average
q8a	%	%	%	%	
0	85.9	69.2	57.5	37.5	
1	14.1	30.8	42.5	62.5	Most Able
q8b	%	%	%	%	
0	89.5	77.9	67.9	41.4	
1	10.5	22.1	32.1	58.6	Most Able
q9	%	%	%	%	
0	97.7	91.5	79.8	52.2	
1	2.3	8.5	20.2	47.8	can't be benchmarked
q10a	%	%	%	%	
0	91.2	72.4	53.1	27.3	
1	8.8	27.6	46.9	72.7	Most Able
q10b	%	%	%	%	
0	92.9	76.1	57.1	31	
1	7.1	23.9	42.9	69	Most Able
q11a	%	%	%	%	
0	95.7	86.9	73.1	52	
1	4.3	13.1	26.9	48	can't be benchmarked
q11b	%	%	%	%	
0	88.5	63.8	39.9	21.4	
1	11.5	36.2	60.1	78.6	Above Average
q12	%	%	%	%	
0	95.7	86.5	73.4	40.7	
1	4.3	13.5	26.6	59.3	Most Able
q13a	%	%	%	%	
0	44.4	14.8	8.3	3.8	
1	55.6	85.2	91.7	96.2	Low Ability
q13b	%	%	%	%	
0	49.2	19.5	11.9	6	
1	50.8	80.5	88.1	94	Low Ability
q13c	%	%	%	%	
0	74.7	41.1	26.8	14.1	
1	25.3	58.9	73.2	85.9	Average
q13d	%	%	%	%	
0	77.3	41.1	25.5	15.8	
1	22.7	58.9	74.5	84.2	Average
q13e	%	%	%	%	
0	90.8	67.6	49.5	26.7	
1	9.2	32.4	50.5	73.3	Above Average
q13f	%	%	%	%	
0	84.6	50.3	28.3	17	
1	15.4	49.7	71.7	83	Above Average

Table 3.11 Item benchmarking bands for Bihar math assessment Std 6

Ability groups	Low Ability	Average	Above Average	Most Able	Benchmarked at which ability group
Table 3.12 Item benchmarking bands for Bihar Math assessment Std 6					
Items	25th Percentile	50th Percentile	75th Percentile	100th Percentile	
q13g					
0	96.6	85.9	70.4	37.9	
1	3.4	14.1	29.6	62.1	Most Able
q13h	%	%	%	%	
0	90.2	62.9	33.3	12.7	
1	9.8	37.1	66.7	87.3	Above Average
q13i	%	%	%	%	
0	96.1	77.6	49.1	22.4	
1	3.9	22.4	50.9	77.6	Above Average
q14	%	%	%	%	
0	78.1	37.8	15.1	6.2	
1	21.9	62.2	84.9	93.8	Average
q15	%	%	%	%	
0	88.3	54.8	26.9	11	
1	11.7	45.2	73.1	89	Above Average
q16	%	%	%	%	
0	95.9	82.7	55.8	22.2	
1	4.1	17.3	44.2	77.8	Most Able
q17	%	%	%	%	
0	98.7	92.2	73.6	37	
1	1.3	7.8	26.4	63	Most Able
q18a	%	%	%	%	
0	96.5	85.3	66.5	27.6	
1	3.5	14.7	33.5	72.4	Most Able
q18b	%	%	%	%	
0	98.2	92.3	79.5	39.4	
1	1.8	7.7	20.5	60.6	Most Able
q18c	%	%	%	%	
0	99	95.9	87.8	60.9	
1	1	4.1	12.2	39.1	Can't be benchmarked
q19a	%	%	%	%	
0	96.4	85.5	64.1	26.9	
1	3.6	14.5	35.9	73.1	Most Able
q19b	%	%	%	%	
0	98	90.6	75.6	40.1	
1	2	9.4	24.4	59.9	Most Able
q20a	%	%	%	%	
0	95.5	81.7	54.7	24	
1	4.5	18.3	45.3	76	Most Able
q20b	%	%	%	%	
0	98.8	95.1	86.8	59	
1	1.2	4.9	13.2	41	Can't be benchmarked
q21a	%	%	%	%	
0	86.4	55.1	24.7	7.7	
1	13.6	44.9	75.3	92.3	Above Average
q21b	%	%	%	%	
0	98	93.6	84.2	59.6	
1	2	6.4	15.8	40.4	Can't be benchmarked
q22a	%	%	%	%	
0	93.8	73.1	42	16.9	
1	6.2	26.9	58	83.1	Above Average
q22b	%	%	%	%	
0	95.3	79.3	49.3	18.5	
1	4.7	20.7	50.7	81.5	Above Average
q22c	%	%	%	%	
0	97.4	89.7	76.3	56	
1	2.6	10.3	23.7	44	Can't be benchmarked
q22d	%	%	%	%	
0	99	95.9	85.6	48.1	
1	1	4.1	14.4	51.9	Most Able

Figure 3.6 Pie chart representation of all math items 6
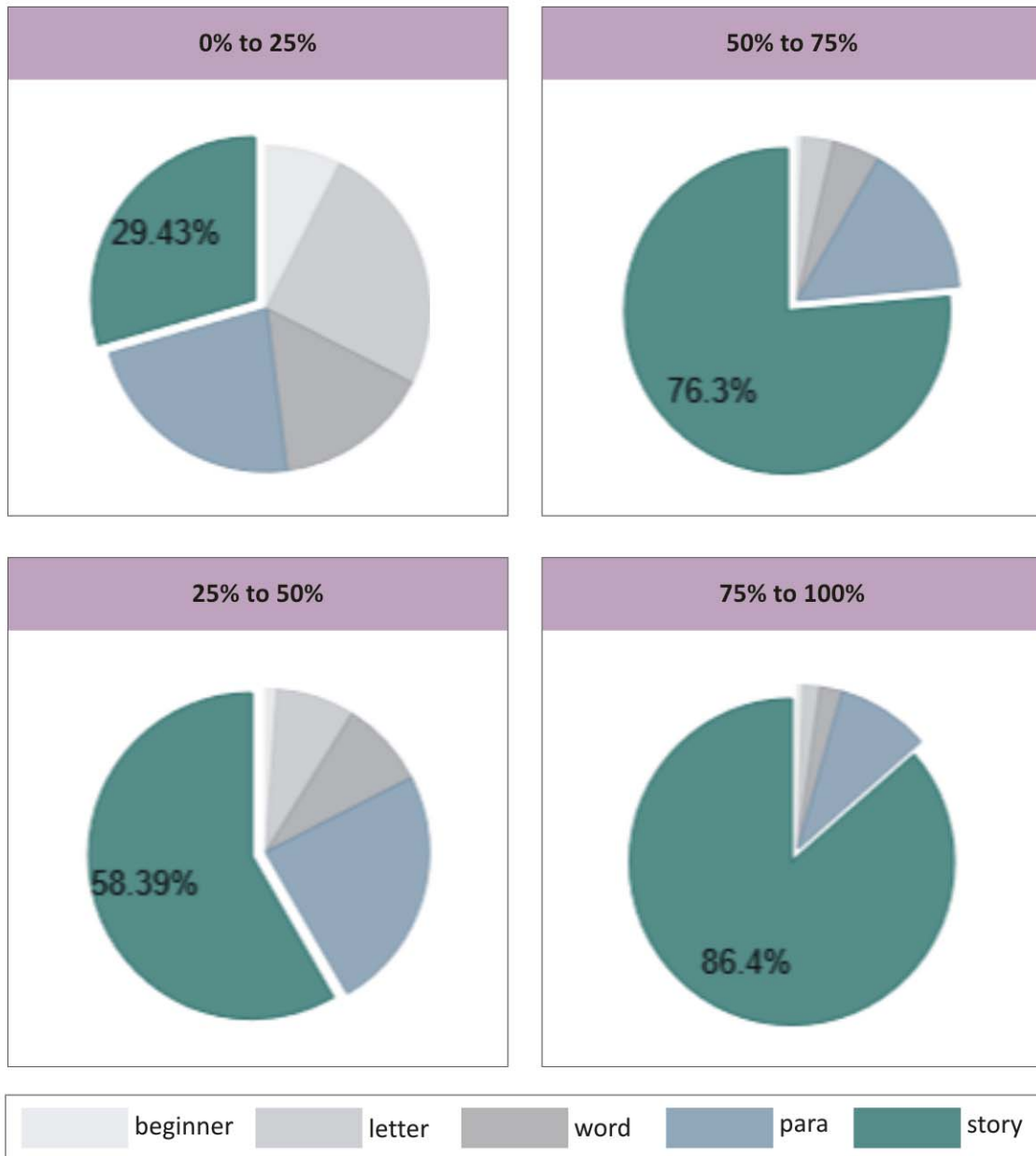
0% to 25%

29.43%

50% to 75%

76.3%

25% to 50%

58.39%

75% to 100%

86.4%

| beginner | letter | word | para | story |

Key Observations

Observe the increasing number of story level students in each of the percentile band from 25th to 100th

Observe the beginner level students who do not know reading. Their number decreases as the percentile band progresses

Observe the letter, word and para level students. Even their numbers decreases as the percentile band progresses Maximum number of story level students are in the top most percentile band

Maximum number of story level students are in the top most percentile band

Conclusive remark (Std 6 Bihar Mathematics Assessment)

The students could on an average achieve less than the ideal mean score. Their ability clusters more on the negative side of the mean and their probability to get full scores decreases with their lowering ability. A very small group of persons who are located on the extreme positive side of the mean have a higher probability to achieve full scores where as the group of people who are located on the extreme negative side of the mean do not have significant or visible probability to attempt the items correctly. The gap between the most able group and the average people is wider than the average and least able group. The test items are pitched at a lower Std level for the students yet they do not find many item easy, most of the items for them are medium to difficult.

Figure 3.7 Item characteristic curves the most difficult item Bihar mathematics assessment Std 6
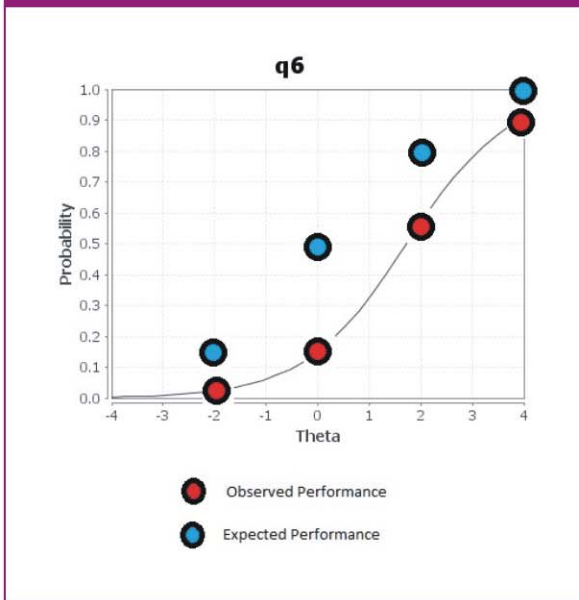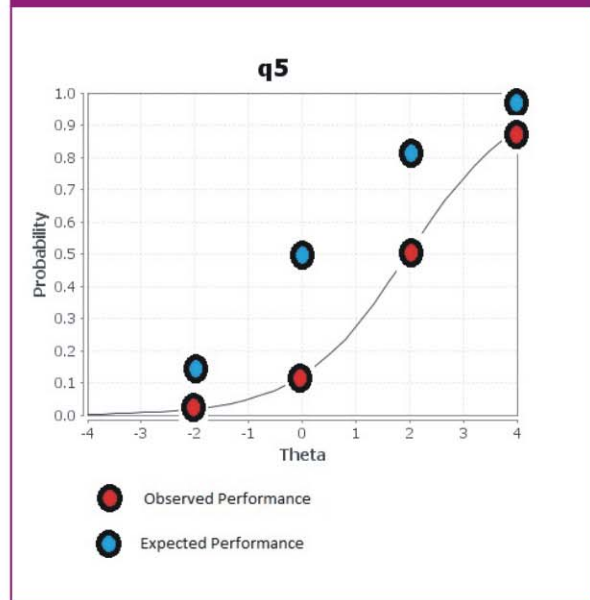

Figure 3.8 Item characteristic curves the most weakly discriminating item Bihar mathematics assessment Std 6

- The overall analysis of the test results and the assessments reveal that even if a assessment is highly reliable and valid, as well as Std appropriate it does not mean that the students will be able to perform appropriately on it.

- The analysis indicates that students uniformly across different levels of performance on reading assessment are not very familiar with the concept of multiple choice items.

- Even if the assessment items have been pitched at a lower Std level still people find it medium to difficult and not easy to deal with. (Example – Language assessment Std 6)

- In mathematics as well, very few items seem easy to the students, rest all seem to be medium to difficult. Whereas the assessment has been pitched at a lower Std level here also.

- Overall students find mathematics test easier than language test. The spread of persons and items on the item person maps and the gap of expected mean versus observed mean also clearly indicate the same in the given figure.

- Reading is one common prerequisite for performing well on any of the written assessments. Here it is clearly evident that those students who cannot read do not perform well on the written assessment as well. These two examples from language and math assessment would support the same. Four items have been taken from both assessment tools and from each set. The point to be noticed is performance on items that require reading and those which do not require reading. The subject could be math or language.

- In both the tests it has been observed that if the student cannot read and comprehend it becomes difficult for him to write or give expression to his thoughts.

- If assessments are meant to gauge the performance of students and their learning levels we find even in the regular constructed response items that there is a gap between being able to read and being able to comprehend or understand it.

Table 4.1 Selected items of Bihar language assessment				
Language - Std 6	Difficulty Level	Set A	Set B	Correct Percentage
Vocabulary	Easy	V2	X	63%
Comprehension Short	Difficult	Cs4 - 18	X	14%
Comprehension Short / Mci	Easy	X	Cs1 - 9	66%
Comprehension Wide	Difficult	X	Cw6 - 18	14%

Figure 4.1 Comparing performances of items that require more reading or less reading Bihar language assessment Std 6

BIHAR LANGUAGE ASSESSMENT - GRADE 6TH

v2   cs4

SET A - MOST EASY ITEM
63% STUDENTS DO IT CORRECT
ITEM LOCATION - -1.14

SET A - MOST DIFFICULT ITEM
14% STUDENTS DO IT CORRECT
ITEM LOCATION - 1.91

Q2. वाक्य पूरा करने के लिए सही शब्द चुनो।

दोनों मित्र देर रात तक बगीचे में ____ रहे थे।
a) टहल ☐
b) खिल ☐
c) चमक ☐
d) उठ ☐
e) उत्तर पता नहीं ☐

PASSAGE WAS GIVEN TO READ BEFORE REFLECTING AND ANSWERING THE ITEM

Q18. अगर मोहन असलम का पक्का मित्र नहीं होता तो क्या वह मोहन की मदद करता? अपने उत्तर का कारण भी बताओ।

उत्तर ....
....
....
....
....

Table 4.2 Selected items of Bihar mathematics assessment				
Math - Std 6	Difficulty Level	Set A	Set B	Correct Percentage
3 digit addition	Easy	13a	X	82%
Read a table and then respond	Difficult	22c	X	9%
3 digit addition	Easy	X	13a	82%
Fractions	Difficult	X	6	10%
Read a table and then respond	Difficult	X	22d	17%

Figure 4.2 Comparing performances of items that require more reading or less reading Bihar mathematics assessment Std 6

BIHAR MATH ASSESSMENT - GRADE 6TH

q13a

q22c

SET A - MOST EASY ITEM
82% STUDENTS DO IT CORRECT
ITEM LOCATION - - 2.8

SET A - MOST DIFFICULT ITEM
9% STUDENTS DO IT CORRECT
ITEM LOCATION - 1.32

Q13. नीचे दिए गए प्रश्नों को हल करो।

a)
```
   1 9 7
 + 8 7 8
 _____
```

Q22. नीचे दी गई सारणी जनगणना की सारणी है। इसमें प्रत्येक गाँव की आबादी की जानकारी है। इस सारणी के आधार पर नीचे दिए गए प्रश्नों का उत्तर दो।

क्र.संख्या	गाँव का नाम	पुरुष	महिला	बच्चे	कुल
1	परवलपुर	400	450	200	1050
2	वासुदेवपुर	300	420	150	870
3	बहादुरपुर	200	210	180	590
4	सबलपुर	700	650	500	1850
कुल		1600	1730	1030	4360

c) किस गाँव की महिलाओं और बच्चों का अंतर सबसे ज्यादा है?

उत्तर यहाँ लिखो

Figure 4.3 Comparing overall performances with in Std 6 across language and mathematics

Bihar Std 6 - Math

Person Plot

Ideal Mean = 20.5/41

Achieved Mean = 16/41

Bihar Std 6 - Language

Person Plot

Ideal mean 9/18

Achieved mean 7.28/18

Bihar Std 6 - Math

Bihar Std 6 - Language
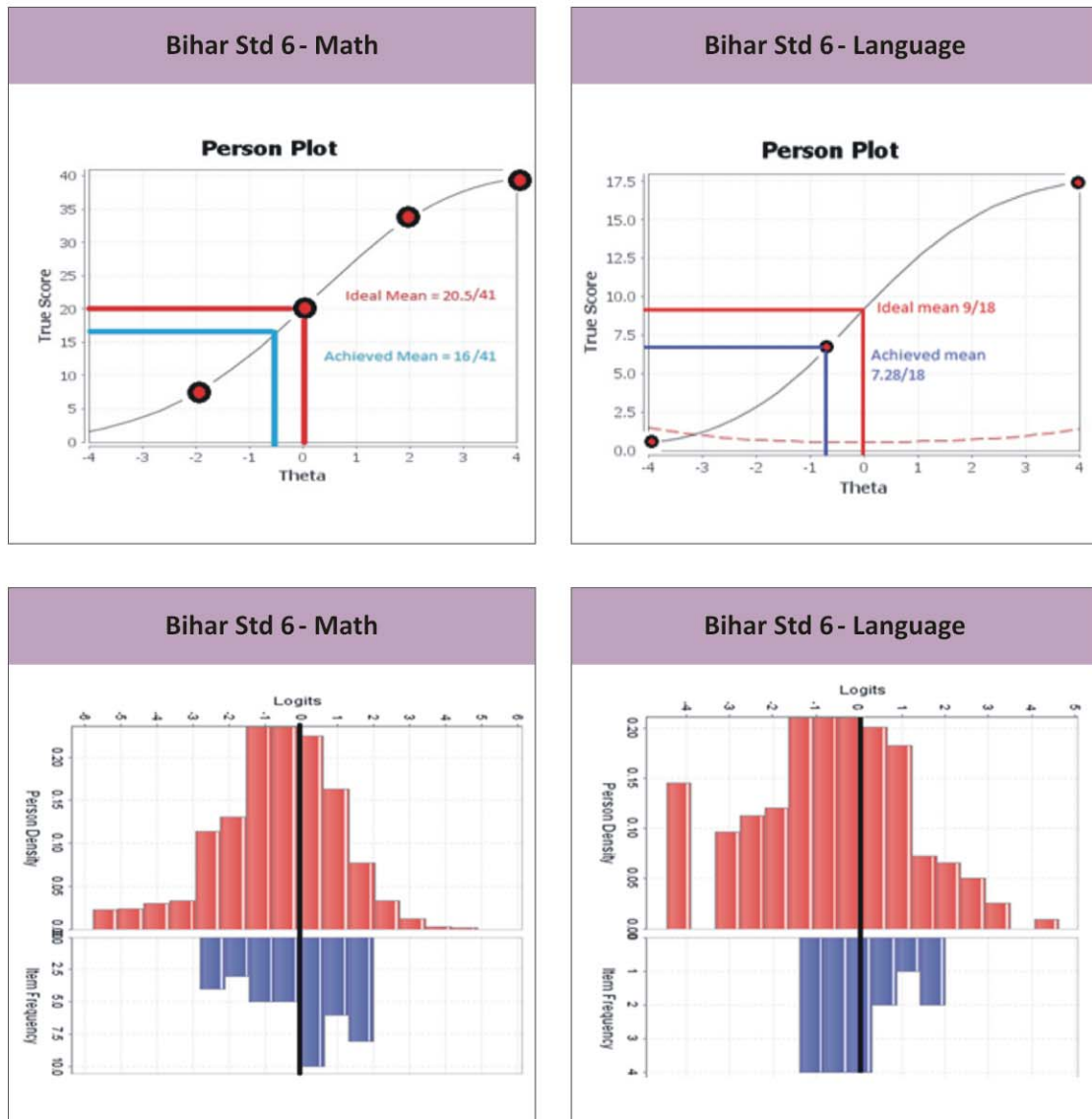
A Modified teaching method which includes frequent communication with the student by adding component of verbal questioning can help in assessing during teaching learning process and giving appropriate feedback.

A model of inter linked teaching learning assessing process needs to be introduced here which has nine components each one of them being interlinked with the other one.

What are we teaching	How are we teaching	Why are we teaching
How are they learning	How are they learning	Why are they learning
What are we assessing	How are we assessing	Why are we assessing
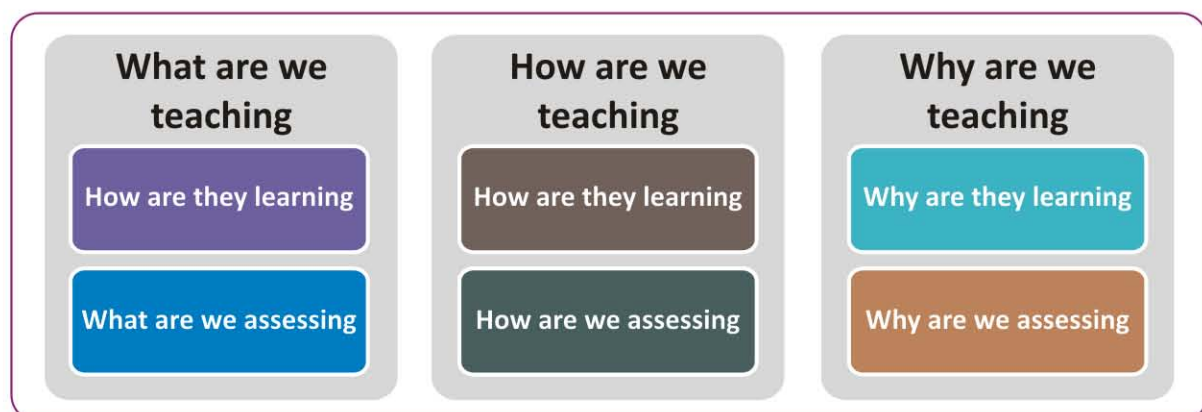
The 'what', 'how' and 'why' of teaching

A teacher usually follows the curriculum and the prescribed books to teach the student so he or she knows what is being taught.

The ways things are taught in a classroom affect the student learning. If more sense organs are utilized to understand a concept the learning is faster and retains for longer. The more a student will see, hear, feel, touch or taste a thing the more he will understand.

Teachers teach in a classroom for various reasons. It is their job or they like teaching or they have to finish the course or they want the student to learn something. The last one is the most important point to ponder upon. Why are we teaching – so that the student can learn something, so that the student can move upwards along the line of developmental continua of any construct or assessable subject matter.

The 'what', 'how' and 'why' of student learning

A student usually follows the prescribed text books and syllabus of the class or Std in which he is studying. Whatever is taught in the classroom is his immediate learning. He might not have a conceptual clarity in the initial level before reading further or before being taught further.

It is the teacher's duty and moral obligation to gauge how he is learning by engaging him in conversation and verbal questioning. The teacher must create an environment so that the student can easily feel good rapport with the teacher and answer the questions to display explicitly his understanding of the concept. Assessment need not always be paper pen or in a strict environment of viva voce. A student will be able to understand where he is right now in terms of learning and where he needs to move up. This can be facilitated by a teacher who can clearly see the difference in students who can read or cannot read, who can read and understand or comprehend and those who can read but not understand or comprehend. The teacher knows what is expected out of the student and how the student can improvise with due guidance and appropriate feedback. This feedback should strictly be about the level of learning and students response. It should not be a very good / well done or very bad / you are foolish type of comment, but rather it could be "you are able to read but you need to understand that..."There could be a situation where the student cannot read at all, he just understands what is told verbally and responds verbally but when told to write he is confused. Every student needs to be treated differently thus. It is most important for a teacher to know at least those set of students in a classroom who do not know how to read. So that first of all they can be taught how to read and measured on all other subjects or assessment levels later on.

Why do students learn – this question again has various answers because they are intelligent, because they are taught, because they need to pass the exam, because growth is inevitable with age and Std etc. But the point to focus here is anything new to a student is a stimulus, if he can just sense it the learning will be short lived, if he is supported to perceive it correctly and understand it then the outcomes will be better. His sensing can be developed to perceiving by the teacher. Reading can be developed to understanding and comprehending by a teacher. Also at what level is the student being questioned matters a lot. Is he being told just to look at some text and recall, recite, pick, match etc then his mind is getting trained to only these kinds of questions. For exams he might just resort to rote learning. But if questions of higher order thinking can be introduced in the classroom teaching learning process then it could be an opportunity for the student to start thinking, to solve problems, to apply their knowledge,

to evaluate, analyze and synthesize or even create. Students need training and grooming to answer higher order thinking items. Teachers can take support from Blooms Taxonomy and its existing set of action verbs to create such verbal items and interact with the students.

The 'what', 'how' and 'why' of assessment

Assessment is a way to measure and gauge student learning. What are we assessing depends on the purpose of assessment. Is it a classroom test? is it a Std level assessment? and is it just a concept clarity assessment? Are we assessing what the student knows or are we assessing what the student does not know are important? Are we assessing what we have taught or are we assessing as per the way we have taught them to respond?

The word assessment does not implicate by any means or ways that it should be paper pen always. It could be just writing anecdotes by observing a student's overall yearlong performance, it could be assessing his homework or a project, it could be a long constructed response or essay, it could be compilation of his everyday responses during the classroom hours, it could be MCI or short answers or long answers.

The most important point of this model is why we assess? Do we assess because we want to see that students have not learnt? or we assess so that we can gauge what students have learnt and not learnt so as to help them to learn? Our motive and a constant motivation to keep this teaching learning process go on matters the most. We all must remember that the end goal is not assessing rather it is improving student learning.

It can be concluded that correct teaching – proper communication with verbal questioning – correct feedback – correct learning – purposeful assessment all of this is interlinked as one cannot work without the other. All the wheels need to spin in same momentum and direction.

How correct are we when we say that the students have performed poorly in exams without knowing whether they can read or not and if so whether they can comprehend or not is a question and a deep thought to be pondered upon.
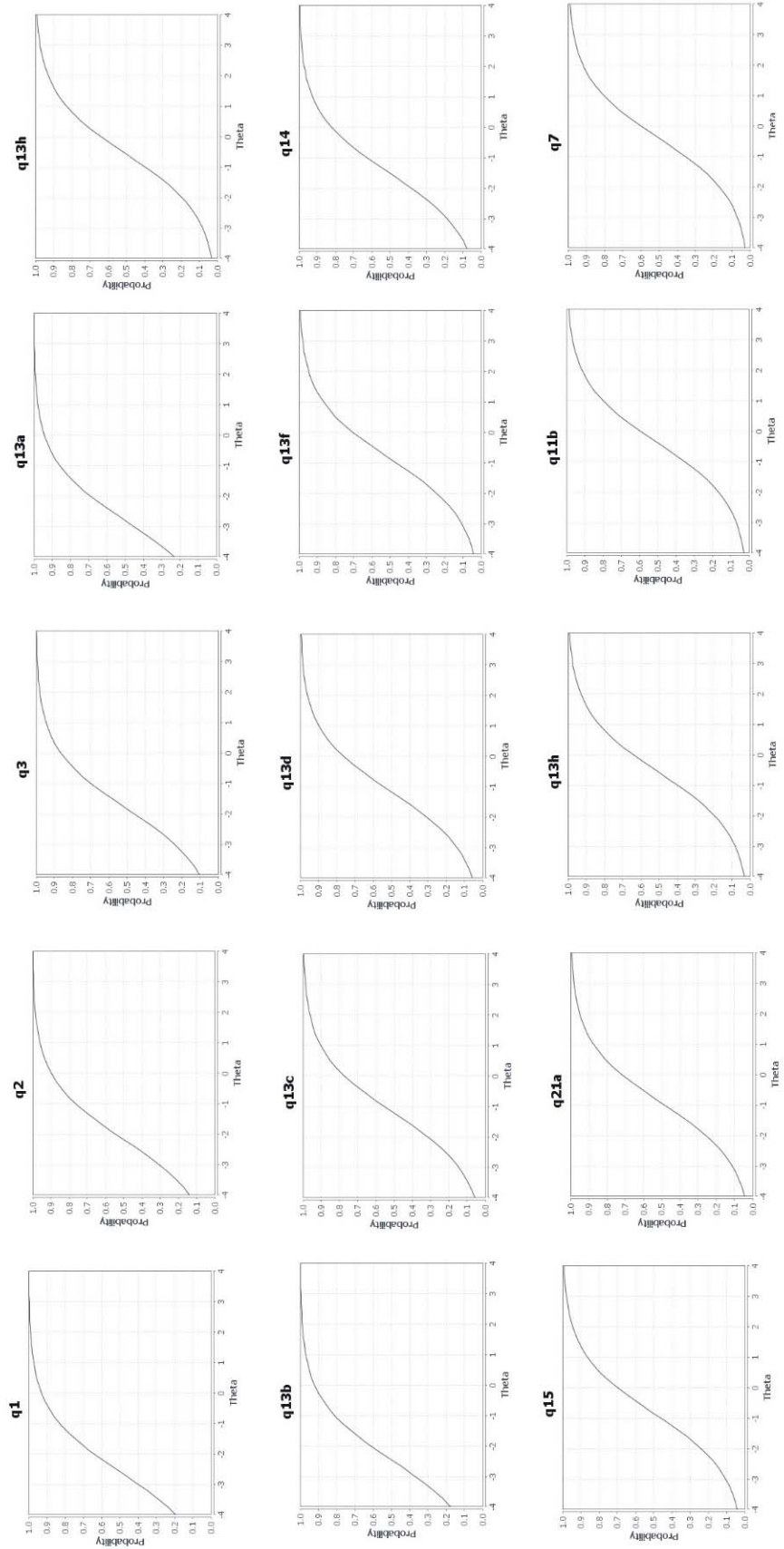
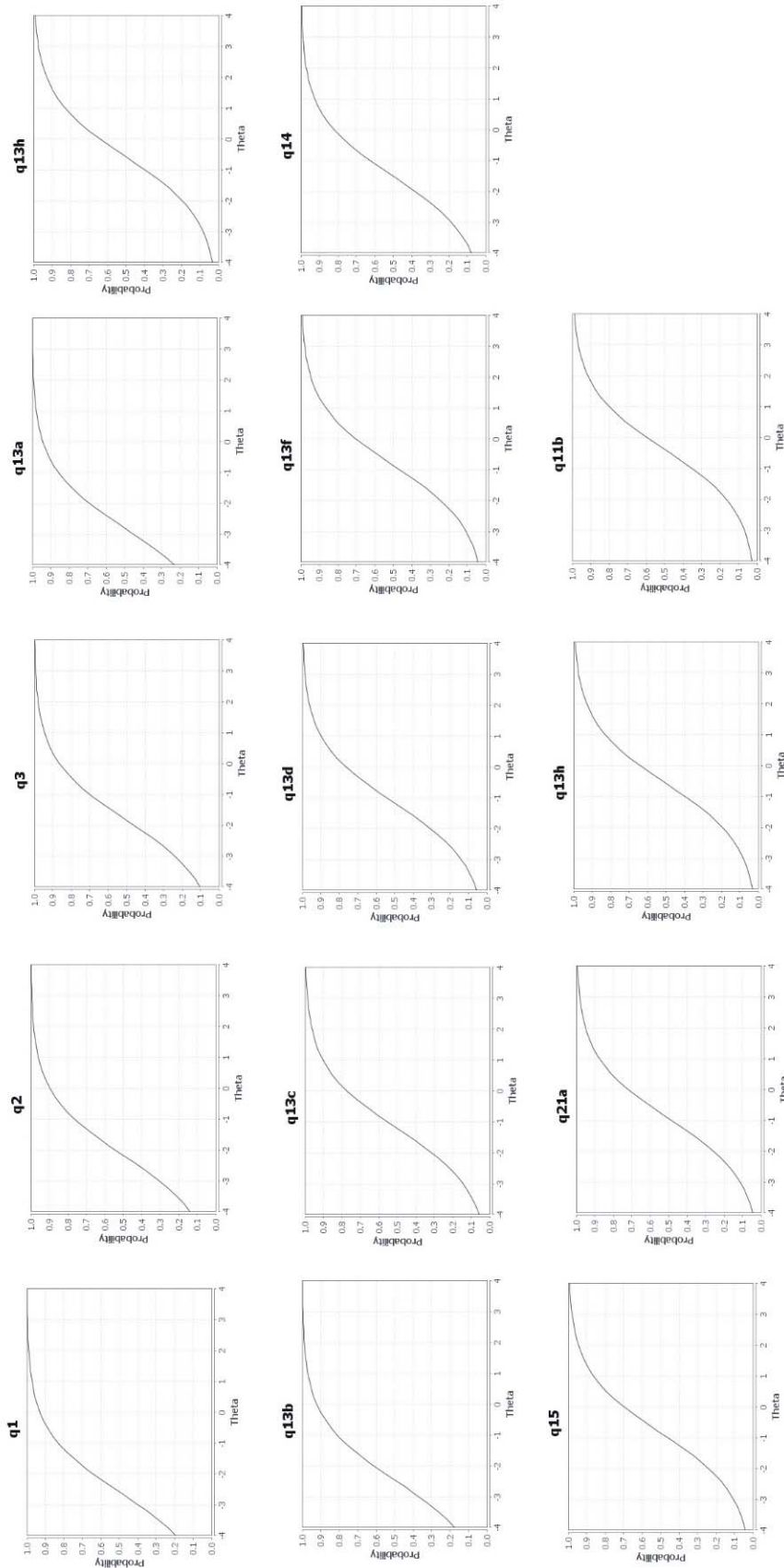## Figure 6.1 Bihar language assessment Std 6
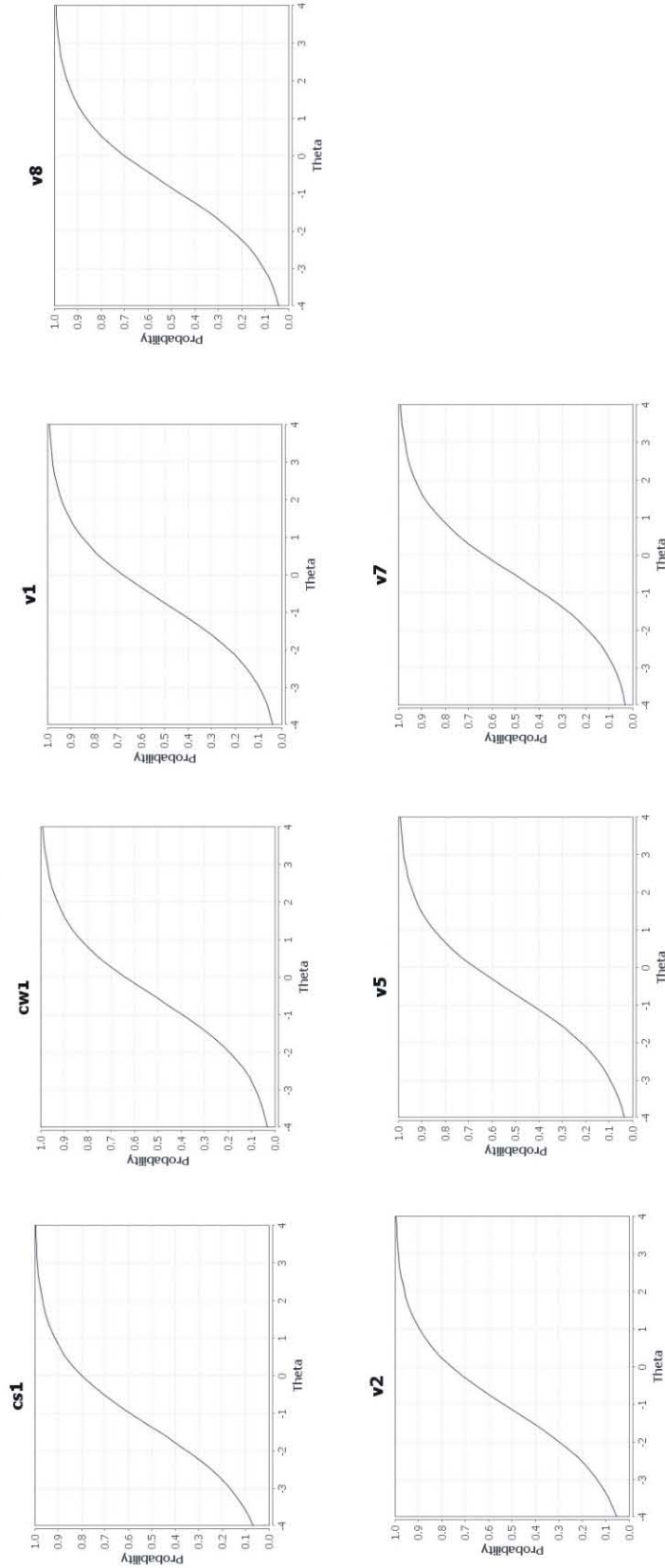
Items in which the students have over performed
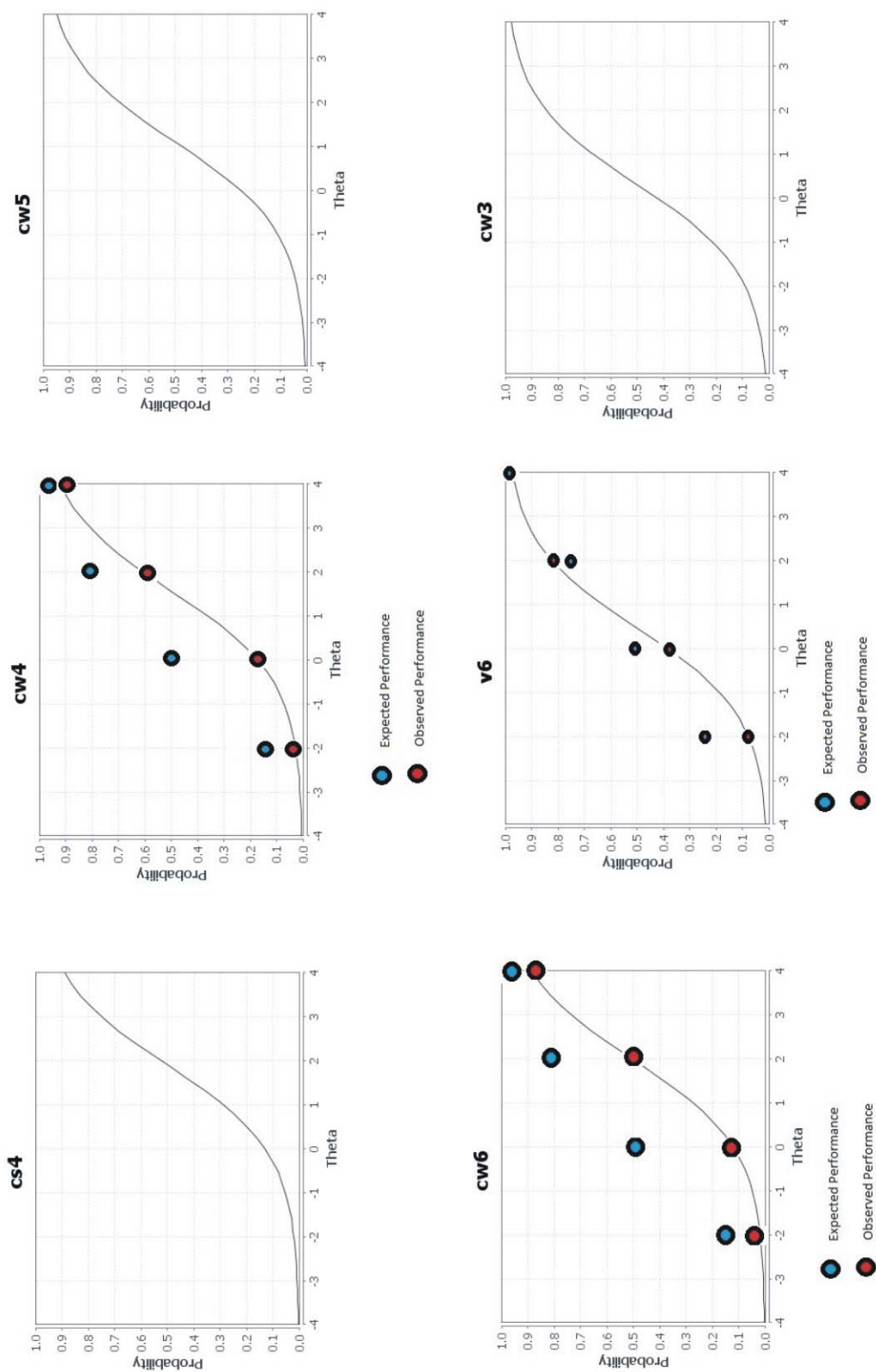
Figure 6.2 Bihar mathematics assessment Std 6

Items in which students have over performed
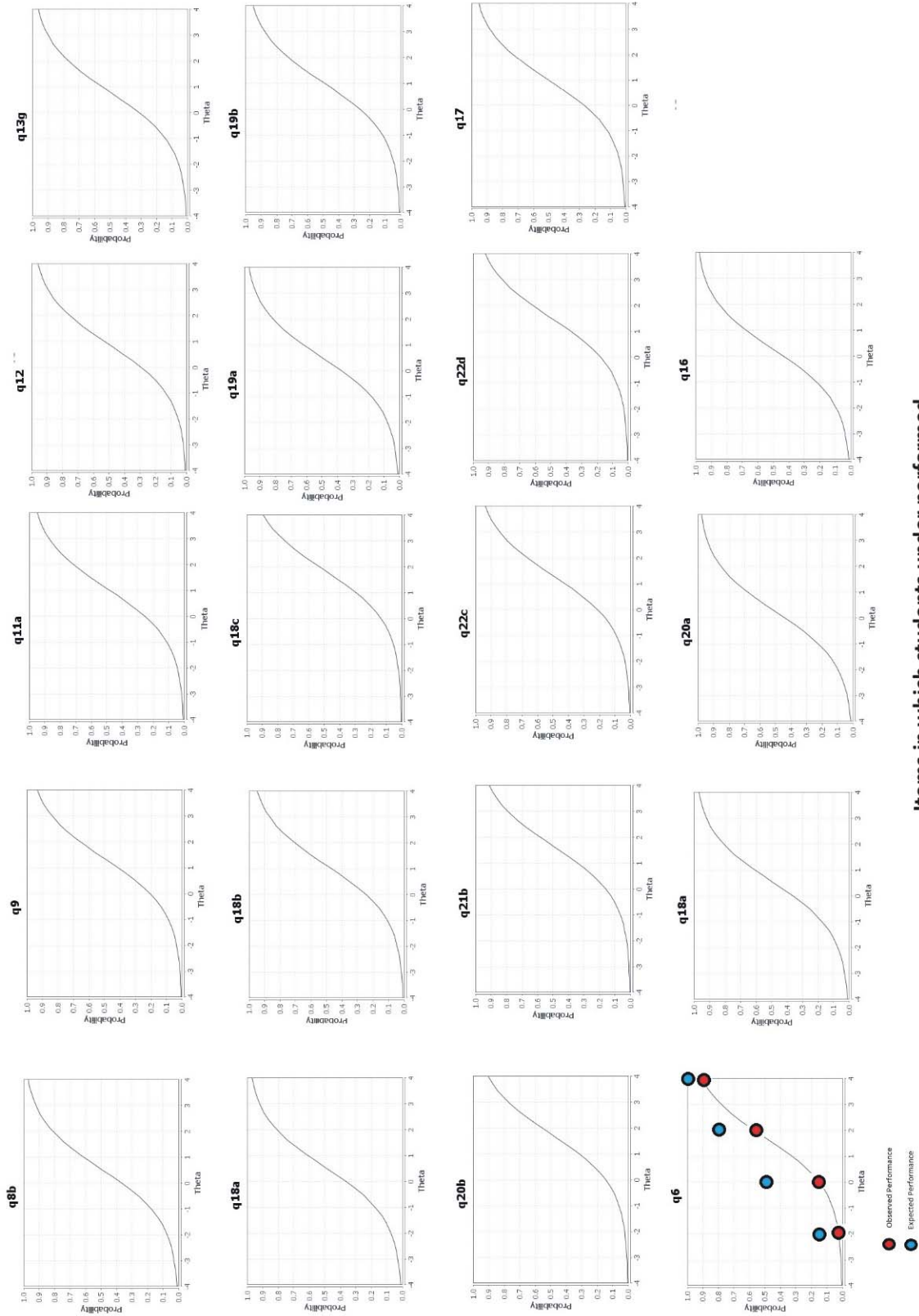
Figure 6.3 Bihar mathematics assessment Std 6

Figure 6.3 Bihar mathematics assessment Std 6

Items in which students have over performed

Figure 6.4 Bihar language assessment Std 6

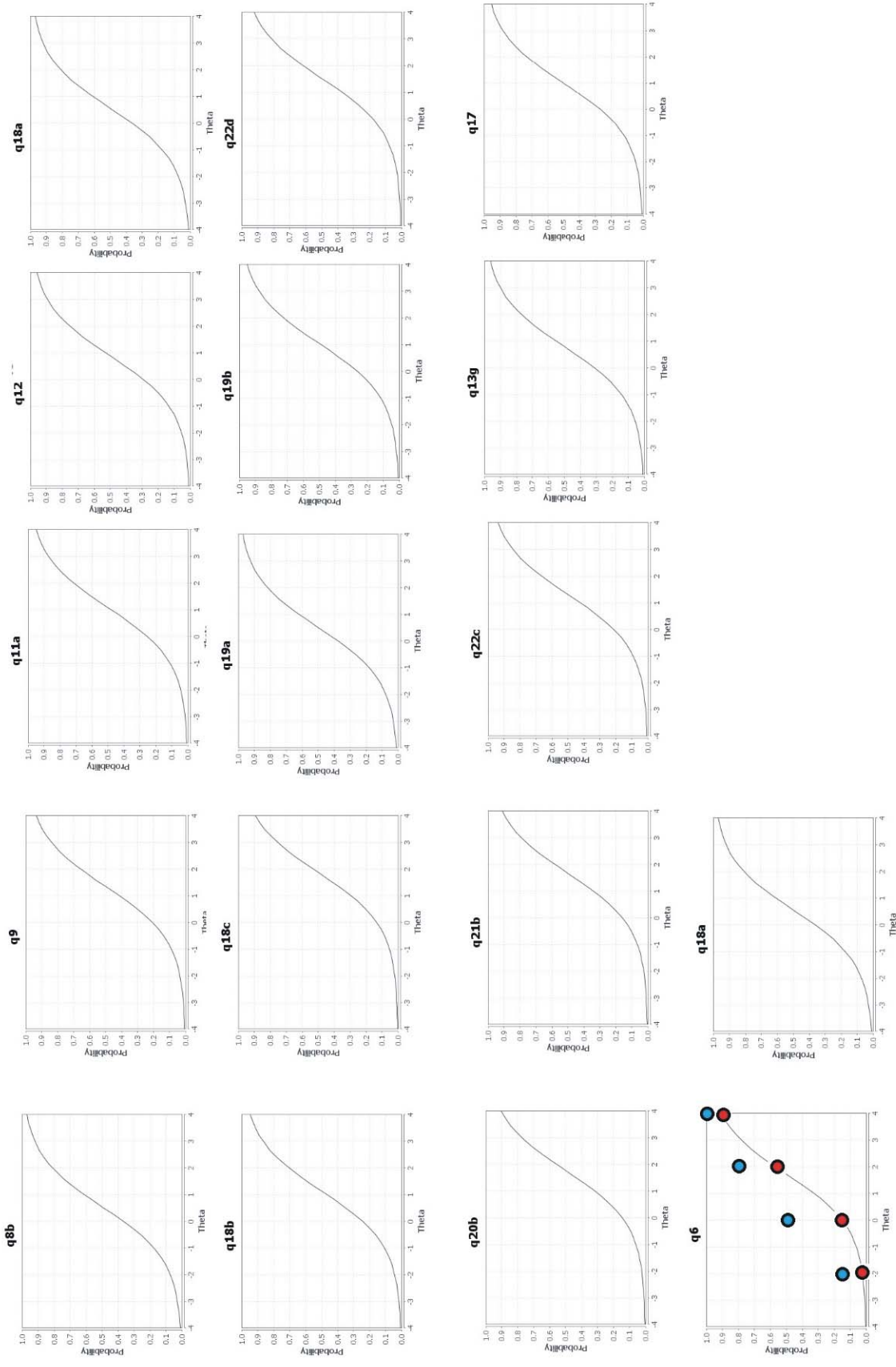**Items in which students under performed**

Figure 6.5 Bihar mathematics assessment Std 6

Items in which students under performed

Figure 6.6 Bihar mathematics assessment Std 6
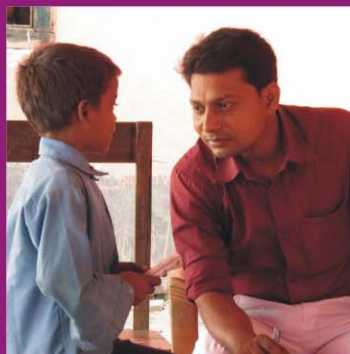
Items in which students under performed

Bihar

School

Report

May 2014