



Education & Literacy Department

Government of Sindh

Sindh Students Assessment

Technical Report

2009

Mathematics Grade 4

Provincial Education Assessment Centre (PEACE)

Bureau of Curriculum and Extension Wing Sindh, Jamshoro

Supported by

Reform Support Unit Sindh, Karachi

May 2010

Acknowledgement

The Education and Literacy Department acknowledges the work undertaken by:

PEACE Sindh, Jamshoro in developing, designing, administering the tests and background questionnaires, analyzing the data and writing the report of the results of Sindh Students Grade 4 Mathematics Assessment 2009.

Focal persons in the districts who managed the distribution of the assessment instruments and the collection of the completed test booklets

Tests administrators who undertook the arduous work of administering the tests and background questionnaires.

The students who took the tests

The Head Teachers, teachers and students who completed the background questionnaires

The RSU who supported the whole activity

EU SER-TA who provided technical support

Acronyms

BoC	Bureau of Curriculum and Extension Wing Sindh, Jamshoro
BQ	Background Questionnaire
EDO	Education District Officer
EU	European Union
FA	Fractions Book A
FB	Fractions Book B
GA	Geometry Book A
GB	Geometry Book B
GEC	Government College of Education
GECE	Government Elementary College of Education
GIS	Geographic Information System
GoS	Government of Sindh
ICC	Item Characteristic Curve
IRT	Item Response Theory
ITEMAN	Item Analysis Program
MA	Measurement Book A
MB	Measurement Book B
MCQ	Multiple Choice Question
MOS	Measure of Size
NA	Number Book A
NB	Number Book B
NEAS	National Education Assessment System
PEACE	Provincial Education Assessment Centre
PITE	Provincial Institute for Teacher Education
PPS	Probability Proportional to Size
PSU	Primary Sampling Unit
QA	Quality Assurance
RSU	Reform Support Unit
SAS	Statistical Analysis System
SEMIS	Sindh Education Information System
SERP	Sindh Education Reform Programme
SER-TA	Sindh Education Reform Technical Assistance

Contents

Page

	Acknowledgement	
	Acronyms	
	List of Tables	
	Executive Summary	
1.	Background	
2.	Introduction	
3.	Sindh Province Testing Model	
4.	Mathematics Assessment 2009 Survey of Grade 4 Students	
5.	Sampling	
6.	Large Scale Testing	
7.	Test Marking and Coding and Data Entry	
8.	Data Cleaning	
9.	Sample Weighting and Estimation	
10.	Data Analysis	
11.	Analysis with IRT Scores	
12.	Constraints, Lessons Learnt and Recommendations	
	Annexes	
	1. PEACE Concept Paper	
	2. Mathematics Test Framework and Specifications	
	3. List of item Writers for Sindh Mathematics Diagnostic Assessment 2009	
	4. Examples of Summary Information of Test Items	
	5. Item Review Checklist	
	6. Sample List for Pilot Testing	
	7. Pilot Tests	
	8. Pilot test Analyses	
	9. Background Questionnaires	
	10. Test Administration Guide	
	11. District Focal Persons	
	12. Large-Scale Tests (Urdu and Sindhi)	
	13. Lead Master Trainers	
	14. District-wise Test Administrators	
	15. Training Centres for Test Administrators	
	16. Marking and Coding Centres	
	17. Example of Coding Sheet (Sukkur)	
	18. Data entry personnel	
	19. Comparison between Census and Survey Enrolment	
	20. Number of PSUs at Sample Allocation, Sample Selection and Data Analysis	

	21. SAS Code to Link Background Questionnaire and Test Data	
	22. Post Stratification Control Totals (Enrolments)	
	23. Results of Data Analysis	
	24. Selected item Characteristic Curves	
	Glossary of Terminology	

List of Tables

1.	National Curriculum Test Specifications	
2.	Pilot Test Items	
3.	Large-scale Test Items	
4.	Sampled Schools and PSUs in Analyses	
5.	Distribution of Test Items – Booklets A and B	
6.	Census 2008/09 vs. PSUs in Analysis	
7.	Distribution of Schools by Range of the Ratio	
8.	Census 2008/09 vs. Survey Enrollment - PSUs	
9.	Modified Census, Census 2008/09 and Survey Enrollment - PSUs	
10.	Number of Students taking both Tests “a” and “B” in the same Area	
11.	Distribution of Students by Number of Tests Taken	
12.	Sample Size (Number of Students) by Test	
13.	Distribution of PSUs by Number of Tests	
14.	Number of Records on the Background Questionnaire Data Files	
15.	Distribution of Students by Background Data	
16.	Sample Size (Number of Students) by Test – Students Background Questionnaires vs. Student Test	
17.	Chi-square Test for Independence of Distribution of Ability Level by Gender	
18.	Chi-square Test for Independence of Distribution of Ability Level by Location	
19.	Chi-square Test for Independence of Distribution of Ability Level by Tests A and B	
20.	Test of Significance of Difference Between Small and Large Class Sizes (Boys and Girls)	
21.	Test of Significance of Difference Between Small and Large Class Sizes (Rural and Urban)	
22.	Estimated Regression Coefficients of the Categories SQ15 – Recoded and Test of Hypothesis that Coefficient is Equal to Zero	
23.	Average IRT Scores for the Categories of SQ38 - Recoded	

Executive Summary

There is general agreement of the need for more consistent efforts to be made to improve learning quality and to measure learning outcomes. The Sindh Education Reform Programme (SERP) aims to do this.

To support improvements in learning quality a mathematics study has been undertaken to provide education decision makers with systematic information about the status of students' learning and the extent to which they attain pre-defined standards and competencies as identified in the 2006 National Curriculum. It enables Sindh province to identify its needs for focused interventions for the improvement of mathematics teaching and students' learning and their learning environment.

The main objectives of this study are summarized below:

- To assess what students in Grade 4 know and can do in mathematics
- To use information regarding students' attitudes to mathematics and mathematics teaching to improve the quality of education and students' learning
- To use information from Head Teachers and teachers regarding their attitudes, school policies and mathematics teaching practice to improve the quality of teacher training and teachers' classroom practice.

The domains of Grade 4 mathematics assessed were number, fractions, measurement and geometry. The survey was conducted with 28,684 students in all districts of the province. Background questionnaires were given to Head Teachers, teachers and students to identify the relationship between, for example, teacher qualifications on student results, giving homework and student results, student attitudes and results.

The main results of the survey are as follows:

- The overall mean score of students was 44.7%;
- Ten out of the 23 districts performed above the mean performance of the rest of the districts in the province;
- Students in rural areas (44.5%) achieved a higher mathematics score than those in urban areas;
- Boys (45.6%) achieved better scores than girls (43.7%)
- Students performed best on number (47.3%) and measurement (47.9%) test items;
- Students obtained the highest scores on procedural knowledge test items (57.7%) followed by conceptual understanding items (53.36%)
- Students were weakest in their achievement of problem solving items (43.8%)
- Students who were taught in groups during most lessons had lower achievement than those who were taught in groups in some lessons, every week;

- Students who were asked to explain the process they used to obtain an answer achieved better scores than those who were not questioned about the process used.

From this study it can be seen that much work is required to improve the teaching and learning of mathematics in the subject areas tested. The results of these tests are found in detail in Section 8.2.

The results of the tests have implications for Sindh Province regarding the need to:

- Identify requirements and strategies and plan for improvements in student learning;
- Interpret the National Curriculum according to the needs of Sindh province;
- Make assessment **for, as** and **of** students' learning central to the development of improved teaching and learning methodologies;
- Develop supplementary materials to support student difficulties and teachers' teaching;
- Improve textbook development in line with the 2006 National Curriculum standards and competencies;
- Improve Teacher Training and Teacher education Development;
- Improve the roles of management in the districts to mentor and advise teachers in a supportive manner.

It is hoped that these implications will be further discussed and integrated into the existing SERP programme.

Lack of quality assurance procedures during the translation of test items, during sampling, test administration and data entry resulted in loss of 13 percent of the PSUs allocated at the sample design stage. In particular, the sample losses were very large for Karachi Rural, Mirpurkhas Urban, Tharparkar Rural and Nawabshah Urban. 83 percent of PSUs from the Karachi Rural stratum, and about one-third of the PSUs from each of the Mirpurkhas Urban, Tharparkar Rural and Nawabshah Urban strata, was lost. This had significant adverse effect on the quality of their estimates. In fact, the CVs of some of the estimates for Karachi Rural were around 50 percent. Therefore, survey estimates for Karachi Rural could not be released because of poor reliability. Moreover, the results of t-tests to test the significance of differences between Rural and Urban have been not reported for Karachi Rural.

The quality of the background questionnaire data also suffered because of not implementing the quality checks. The background data were available from all three questionnaires (i.e., Head Teacher, Teacher and Student) for just over half (51.7 percent) of the students. The loss of background questionnaire data were the result of either not completing the questionnaires or providing erroneous IDs resulting in no-matched background questionnaires. Moreover, the questionnaires that matched contained large amount of missing and/ or invalid data. As a result, no analysis could be conducted by combining data from all three questionnaires (i.e., Head Teacher, Teacher and Student).

PEACE in future testing needs to ensure that quality assurance procedures for all aspects of testing are in place to ensure the quality not only of the tests but test administration, data entry and background questionnaire information.

1. Background

Education is considered to be a major and important factor in the overall development of the country. The Government of Pakistan has recognized this and is a co-signature of the Dakar Framework for Action, 2000 which identifies two goals related to improvements in the quality of education, namely:

Goal 2: Ensuring that by 2015 all children, particularly girls, children in difficult circumstances and those belonging to ethnic minorities have access to complete free and compulsory primary education of **good quality**.

Goal 6: Improving **all aspects of the quality of education** and ensuring excellence of all...

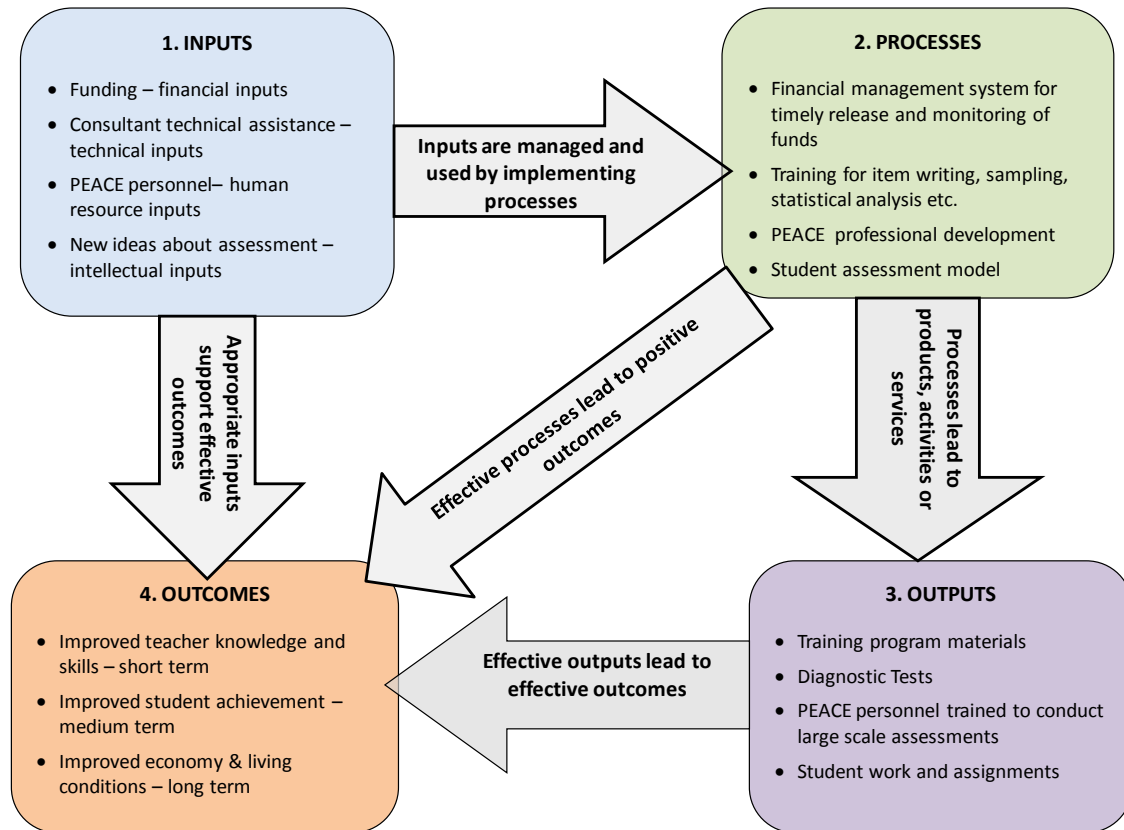
Sindh Province has recognized that education is important at all levels and is particularly emphasizing the need for improvements in education particularly at primary and elementary levels. The Government of Sindh is committed to improving the quality of education through the Sindh Education Reform Programme (SERP).

Quality in education requires:

- Standards of achievement to be agreed, set and used
- The creation of a foundation of teaching methods that work and are used in the classroom, and
- The creation of a system of accountability to measure the results to identify areas of strengths and weaknesses to enable the identified improvements required to be made to be targeted.

SERP has identified the need for standardized assessments to be undertaken to provide more detailed, disaggregated information regarding student learning in each district of the province. These assessments will provide detailed information for use in the development of improvements in delivery of the curriculum, improvements in teacher training, textbook development and create a system of accountability.

The Outcomes in the SERP are identified as improved teacher knowledge and skills, improved student literacy and in the long run an improved economy and living conditions for all in Sindh Province. The inputs, outputs and processes identified by SERP to achieve these outcomes are found in the diagram below.



2. Introduction

2.1 The Role of PEACE, Sindh Province

PEACE, Sindh was initially developed as a unit to support the development of national assessments conducted by NEAS. Between 2004 and 2008 four assessment surveys were conducted in languages, mathematics, science and social studies. PEACE personnel obtained capacity building in item writing and test development, test administration and training of test administrators, verifying the NEAS sample, identifying policy issues for background questionnaire development, marking and scoring and statistical analysis and report writing and the dissemination of results through this development.

PEACE has now established its own role through the development of provincial assessments aimed at identifying in each district the strengths and weaknesses of the achievements of students in relation to the National Curriculum, and correspondingly the strengths and weaknesses of the teaching process and textbooks used in the classroom. PEACE proposals and the rationale are found in a Concept Paper developed in 2007 (Annex 1).

2.2 Assessments conducted in Sindh Province

At present assessments in Sindh Province, with the exception of the National Education Assessment System (NEAS) assessments are conducted in a non-standardized manner. End of semester examinations are non-standardized. They are not developed according to agreed standards of student achievement and are not administered or analyzed in a standardized manner. This does not enable comparisons of achievement levels to be made in the province on a year to year basis or comparisons to be made between schools, districts. It does not enable credible data to be provided to the education system for informed policy decision making including the targeting of financial, technical, human resources and intellectual inputs in a more effective, efficient, relevant manner to ensure that educational inputs have the desired impact on the quality of education.

2.3 Proposed Three Year Action Plan for PEACE

To enable Sindh Province to judge the quality of education it is necessary to obtain evidence of the achievements of students against agreed standards. This requires evaluating the extent to which students have developed their knowledge, understanding and skills in specific subject areas.

PEACE has proposed a cycle of testing to be conducted over the next three years. This involves testing the four core subject areas of mathematics, language, science and social studies according to the following timetable.

2009	2010	2011	2012
Grade 4 Mathematics	Grade 4 Languages	Grade 4 Science Grade 8 Mathematics	Grade 4 Social Studies Grade 8 Languages

2.4 PEACE Tests in The Context of The Nature of Assessment

Assessment has many purposes. It is about:

- Reporting on students' achievements;
- Improving their teaching through expressing more clearly the curriculum goals;
- Measuring student learning;
- Diagnosing misunderstandings in order to help students to learn more effectively;
- The quality of teaching as well as the quality of learning.

Assessment is central to the learning process and is a crucial aspect of teaching. Research over time has identified it is the most significant factor that influences student learning.

If we wish to discover the truth about an educational system, we must look into assessment procedures.... The spirit and style of student assessment defines the defacto curriculum. (Rowntree, 1987, p1) ¹

Assessment methods and requirements probably have a greater influence on how and what students learn than any other single factor.....This influence may well be of greater importance than the impact of teaching materials (Boud 1988)²

What influenced students most was not the teaching but the assessment (Gibbs and Simpson, 2004, p4)³

In the light of research information, the NEAS test results and the need for improvements in teaching and learning and defining what needs to be included in textbooks and the locally based curriculum, PEACE has developed a series of diagnostic tests in mathematics. These diagnostic tests will identify where students are having particular difficulties in the specific areas tested. An example of test items to test more progressively difficult concepts is given in Box 1 below, Addition of Numbers Using Vertical Form.

Box 1: Addition of Numbers Using Vertical Form.

(a)	(b)	(c)	(d)
5	7	12	45
+ <u>2</u>	+ <u>9</u>	+ <u>22</u>	+ <u>17</u>

This requires the student to demonstrate an understanding of (a) addition of units without carrying; (b) addition of units with carrying ten units (c) addition of tens and units without carrying; (d) addition of tens and units with carrying.

Information obtained from the analyses of these test items will enable teachers and student to focus on those areas that require further teaching and also provide information to teacher trainers, curriculum and textbook developers regarding the areas they should focus on.

3 Sindh Province Testing Model

The testing model used for the Sindh Provincial Assessment is based on the need to identify student achievement. This is different from the limiting psychometric model which emphasizes ranking and statistically derived distributions since it involves a shift away from a norm referenced approach towards one where what students can do is stated. This requires descriptions of performance as found in the Mathematics 2006 National Curriculum for Pakistan.

¹ Rowntree, D. (1987) *Assessing Students – how shall we know them?* London Harper, and Row

² Boud, D. (Ed) (1988) *Developing Student Autonomy in Learning*, 2nd Edition, London, Kogan Page

³ Gibbs, G. and Simpson, C. (2004/05) *Developing Conditions under which Measurement Supports Students' Learning: Learning and Teaching in Higher Education* 1, 3-31

The model also looks at the different levels of achievement of students according to the requirements of the Grades 1 – 4, National Curriculum.

The use of performance descriptions has implications for reporting the results. While the use of a single overall figure gives us some notion of student achievement there is also a need to provide qualitative descriptors what students can do according to the areas tested and by denoting the levels attained within the subject areas assessed.

This model was therefore developed through analyzing and mapping the Grade 4 Mathematics, 2006 National Curriculum and identifying the competencies to be tested. From this two domains were identified for testing, namely, context and cognitive domains and test specifications were written to develop test items to assess these specific domains according to the weightage indicated in the National Curriculum.

Test items were written and classical item analysis (ITEMAN) was used to identify item difficulty and the ability of each item to discriminate between students of different abilities.

Following this, test items were identified for use in the large-scale testing. A two parameter model was used in the final data analysis to identify item difficulty and the ability of each item to discriminate between students of different abilities. This provided results according to, for example, location, gender, districts. A regression analysis (linear/logistic) was used to identify such aspects as the impact of teachers on learning.

The sampling model used was a two parameter model based on district (23 districts) by location. This resulted in 46 strata (e.g. rural, urban) for the province.

4. The 2009 Mathematics Survey of Grade 4 Students

This information is only related to academic achievement and the effect of different background variables and attitudes on student achievement. The privacy and therefore the identity of individual students and families as well as the identities of the participating schools are not released.

The Grade 4 Mathematics Assessment, 2009 is a provincially representative assessment of what Sindh Province students know and can do in various aspects of Mathematics. The first Mathematics Provincial Assessment was conducted in all the districts of Sindh in 4333 schools (Primary Sample Units) and with 28,684 students who answered 106,716 tests.

4.1 Test Framework Development

The development of a test framework and specifications are essential if the testing is going to measure the elements for which it has been constructed. A test has to have a clearly stated purpose and should clearly describe the content and cognitive domains for the grade for which it has been developed. Also the length of time of the test should also be determined as this will have a direct effect on the number of items in the test and also the breadth of the curriculum to be tested.

PEACE along with working school teachers, BoC, PITE and GECE staff developed a Mathematics Assessment Framework in 2008 (Annex 2). This framework identified the two dimensions of mathematics in the National Curriculum to be tested, namely subject content domain (defines the subject matter covered by the assessment) and the cognitive domain (knowing facts and procedures, using concepts, solving problems, reasoning). This was the foundation of the provincial assessment and the basis for item development. The content domain areas to be tested were identified in the Assessment Framework as number, fractions, measurement and geometry. Each content domain has several topic areas (e.g., number is further categorized by whole numbers, integers, and ratio, proportion, and percent etc.).

The development of the Mathematics Assessment Framework included the development of a test specification. This described the number of questions and the weightage of each area to be tested. This ensured that the content domains of these areas were more likely to be assessed in a balanced way.

The writing of the test specification required:

- Using the Mathematics 2006 National Curriculum and the results of the NEAS tests to identify the content areas to be tested;
- Identifying from the National Curriculum the weightage to be given to each area to be tested;
- The cognitive areas to be tested (conceptual understanding, procedural knowledge and problem solving)

The National Curriculum Test Specifications for each of the curriculum content areas were identified as follows:

Table 1: National Curriculum Test Specifications

Mathematical Abilities	National Curriculum Content Areas			
		Numbers (%)	Fractions (%)	Measurement (%)
Conceptual Understanding	40	40	41	50
Procedural Knowledge	45	53	43	45
Problem Solving	15	7	16	5
Total	100	100	100	100

4.2 Review of the 2006 National Curriculum and Mapping the Mathematics Curriculum

Before item writing, a workshop was conducted regarding the National Curriculum and the mapping of the mathematics curriculum. As this was a diagnostic test it was important for all to

understand the mathematics curriculum and its requirements and also mathematical progression between the grades. Curriculum mapping aligns assessment and the curriculum. It is a technique for:

- Exploring the primary elements of the curriculum – what is taught, how instruction occurs, when instruction is delivered;
- Identifying seams and gaps;
- Identifying repetition between scope and sequence;
- Allow vertical alignment of assessments, content and methods across grades;
- Support horizontal alignment of assessments, content and methods;
- Improve assessment

The curriculum mapping that was undertaken consisted of both horizontal and vertical mapping. The documentation is found in Annex 2.

4.3 Item Writing

Test Item Workshops were conducted in 2008.

Using the Test Framework and Specifications test items were written. All the items were multiple choice items (MCQs) as it was agreed that these would be more likely to provide objective results and would enable easier standardization of test setting and marking.

Items were written according to the following instructions:

- Items should match the objective to be tested
 - ✓ The test items should only ask students questions they should be able to answer
- Items should test at the appropriate cognitive level (the type of thinking involved)
 - ✓ What? (Knowledge and Concepts)
 - ✓ How? (Procedures and Performances)
 - ✓ Why? (Problem Solving and Reasoning)
 - ✓ Objectives should contain action verbs.
 - ✓ Test items should use action verbs.
- Items should be written in simple language for Grade 4 students and should be correct and clear

When the items were written the item writers (Annex 3) were required to identify the competency they were addressing, the estimated ease of the item and also why they had chosen the distractors for the test items. An example of the flimsie is found in Annex 4.

Seven hundred and seventeen test items were written for all four areas. An item review process was undertaken through meetings and a final review was completed by the PEACE. The review process was supported by a checklist (Annex 5)

The final number of test items accepted for pilot testing was as follows:

Table 2: Pilot Test Items

Mathematical Abilities	Curriculum Content Areas				
		Numbers	Fractions	Measurement	Geometry
	Conceptual Understanding	55	59	156	98
	Procedural Knowledge	50	67	122	55
	Problem Solving	7	13	25	10
	Total Number of Items	112	139	303	163

4.4 Pilot Testing

Pilot testing was conducted in November 2009. Pilot testing was required for the PEACE to ensure that the demands of the tests were appropriate and also to identify items which were reliable, valid and discriminated appropriately between the different abilities of students.

The pilot tests (Annex 7) were developed from the pool of items identified in Table 2 above. Two tests, of 50 test items each for each area to be tested were designed. Items which tested the key competencies in the identified mathematical areas and those competencies that were able to be tested in a pencil and paper test were to be included. The weightage given to the specific content areas was according to the weightage given in the 2006 National Curriculum.

For the piloting, 46 schools were selected from 23 districts (2 schools from each district) A representative sample of 1150 students from 46 schools (34 Sindhi Medium and 12 Urdu Medium Schools) representative of rural and urban areas and male and female students in the province took part in the pilot testing in November 2008. A list of the schools participating in the pilot tests is found in Annex 6.

4.5 Analysis of Pilot Test Results

The pilot tests were coded and entered into a database. An analysis of the pilot test results was conducted to identify “good” test items. The items were Classical as “good” if they demonstrated good reliability (the likelihood of the results to be replicated on subsequent occasions) and validity (whether the knowledge, skills which the test is supposed to measure are being measured) and if they were able to discriminate between students of different abilities. Item analysis of the pilot test items was carried out using **ITEMAN** software. The results of the ITEMAN analysis is found in Annex 8.

4.6 Retention of Test Items for Large Scale Testing

Using the results of the item analysis of pilot tests and subject specialists’ professional judgment the following number of test items were retained for use in the large-scale testing.

Table 3: Large-Scale Test Items

Mathematical Abilities	Curriculum Content Areas				
		Number (%)	Fractions (%)	Measurement (%)	Geometry (%)
	Conceptual Understanding	36	30	35	47
	Procedural Knowledge	49	50	45	42
	Problem Solving	15	20	20	11
Total % of Items	100	100	100	100	

These test items were organized into two test booklets for each subject content area to be tested. This would ensure that the entire curriculum to be tested was covered and also would ensure as far as possible that little malpractice would take place. Due to the large numbers of items required for each area tested to provide diagnostic information, it was not possible for parallel booklets to be developed.

4.7 Development of Background Questionnaires

Background Questionnaires for Head Teachers, Teachers and students were also developed and piloted (Annex 9). These questionnaires looked at such things as school conditions and environment; teachers and teaching practices; supporting inputs for schools; and, students' home backgrounds.

Some of the difficulties identified in test and Background Questionnaire development are listed below:

- The timeframe for the development of a full-scale mathematics survey in 2009 was extremely tight and required activities to be conducted in a timely manner according to the action plan.
- Many of the item writers lacked awareness of the 2006 Mathematics National Curriculum, the difference between the curriculum and textbooks, a lack of understanding of the different dimensions of mathematics in the form of content and cognitive domains; as well as a lack of understanding of the curriculum's terminology, for example, standards, competencies, benchmarking.
- Ensuring that there were sufficient items to cover the curriculum to be tested as identified in the Test Specifications. While it was expected that only 100 test items were required in total for two test booklets (50 items for each test booklet), 400 items in total, and 707 test items were produced for piloting, the discard rate after piloting was high and all the items identified as "good" were retained.
- Difficulties were found in constructing some of the questions, as well as ensuring sufficient coverage of each background and context variable in relation to the length of the questionnaires and the time it would take personnel in the schools to complete them. Other difficulties occurred in the translation of the Background Questionnaires into Urdu and Sindhi and some of the translated questions proved to be ambiguous. For example, a question written in English: What language(s) do you mostly use at home to talk to

members of your family?, when translated to Urdu and Sindhi became Do you speak your mother tongue at home?

4.8 Development of Test Administrator Manuals

PEACE developed Test Administrator Guidelines based on the NEAS test administration guidelines as the basis. All the Guidelines were provided in Urdu and Sindhi. The Guidelines guided the test administrators using a step by step procedure. The manual was piloted during the piloting of the pilot tests and amendments were made as required. An example of the guidelines is found in Annex 10.

The following aspects were included in the Test Administrator Manuals. It identifies what a test administrator should do before testing, during and after testing and also contains information regarding the file for test administrators and the need for security of the tests.

A. Before Testing

1. Delivery of Assessment Materials to the school
2. Preparation for Assessment Activities in the school
3. Selection of Students to be assessed
4. Filling of attendance sheet
5. Opening of Sealed Envelope
6. Head Teacher Questionnaire

B. During Testing

- a. Preparing of Test Room
- b. Distribution of Assessment Materials
- c. Use of Assessment Booklets
- d. Sample and Exercise Question
- e. Duration of Test
- f. Instruction During Test

C. After Testing

1. End of Assessment
2. Questionnaire for students and parents
3. Arrangement of Assessment Materials
4. Return of Assessment Materials

D. File for Test Administrators

All research/assessment is conducted on the bases of empirical evidence. Therefore to collect the necessary information from the field, different documents have been prepared to make the data entry process easier. The Test Administrators are asked to send these files to PEACE after the completion of assessment activities in the schools .The file contains

1. List of Assessment Materials
2. Attendance Sheet

3. Field Report for Test Administrators
4. List of School Facilities
5. Record of use and unused booklets
6. Checklist for Test Administrators
7. Random Number Selection Form

E. Security of Test Materials

Security of the assessment material is required to;

- protect the test items from being accessed by the public so that they can be used in the future; and,
- ensure the completed tests, questionnaires, TA files are not available to public view – they are kept private.

The security of assessment material is assured by sealing it in envelopes and packing the envelopes in the boxes provided, to prevent any damage. The sealed assessment material is sent to the district focal persons where the assessment material is given to the TAs after the Test Administrators Training Workshop. The sealed material is required to be opened on the day of testing in the presence of the Head Teacher or the In-charge teacher. Monitoring Reports also contain questions on the security and secrecy of assessment material.

A similar procedure is undertaken for the return of the assessment materials

5. Sampling

5.1 Why sample?

Sampling facilitates the assessment process when a large number of students are involved and it is not feasible to assess all the students. In Sindh Province there were approximately 447, 000 students⁴ enrolled in Grade 4 during the 2008/2009 academic year. To test all of these students would not be feasible because of time and budget constraints.

5.2 Sample Design

A stratified two-stage sample design was used for selecting the sample of students who were enrolled in Grade 4 in the government schools in the province of Sindh during the 2008/2009 academic year. The objective of the study was to conduct analyses simultaneously for the Province, location type (Rural/Urban), boys/girls within the province, and the 23 Districts in the province. Therefore, the stratification was defined by cross-classification of District by location type (Rural/Urban) resulting in 46 design strata. At the first stage, schools (or clusters of schools) with Grade 4 classes were selected with Probability Proportional to Size (PPS) systematic sampling, and at the second stage students which were the ultimate sampling units were selected with systematic sampling.

⁴ Source: 2008/09 Annual Census of Schools

The measure of size (MOS) for the PPS sampling of schools was the Grade 3 enrolment from the 2007/08 Census of Schools. The Grade 3 enrolment was used as measure of size because the current Grade 4 students were in Grade 3 at the time of the 2007/08 Census of Schools. Since the 2007/08 Census data was only one year old, good correlation between the current Grade 4 enrolment and the MOS from the Census was expected.

All government schools in the province of Sindh with Grade 4 classes during 2008/09 were part of the target population. Although, the *desired target population* was the population comprising all Grade 4 students in the government schools in the province, it was not cost effective to sample very small schools, in particular in the rural areas where travel costs are very high. Therefore, very small schools although in the target population were consciously excluded. The remaining schools formed the *survey population*. Exclusions were kept to a minimum and used as a means to reduce cost while still selecting as close as possible to a representative sample. International studies have routinely set the upper limit of exclusions at 5.0 percent of the desired target population. Since the analysis was needed by rural/urban schools, the cut-off value used the criteria that the percentage of students excluded from the survey was less than 5 percent both in rural and urban schools. The survey population was therefore all government schools in the province for which Grade 3 enrolment from the 2007/08 Census of schools was greater than or equal to 4 students. The survey population was approximately 447,000 students with almost 70 percent in the rural schools.

As required, the analysis from the 2008/09 survey of Grade 4 students was conducted at the Province, location type (Rural/Urban) within the province, and District levels. The sample of schools was allocated to the 46 design strata defined by cross-classification of the 23 districts and location type (Rural/Urban). In order to conduct analyses for the Province, Rural/Urban type and the 23 Districts the sample size was roughly 4,000 schools (PSUs) with 10 students selected per sampled school resulting in an overall sample of about 40,000 students. The above sample size was arrived at by using the criteria that there was a need for an effective sample of about 300 students for the smallest district. The effective sample size is defined as the actual number of students selected divided by the design effect⁵. The typical design effect for education studies is around 3 to 4. Therefore, the actual sample for the smallest district would be more than 1,000 students. The sample allocation to the two stages of sampling was determined on the basis of cost and variance consideration. First, the total sample was allocated to the 46 strata defined by cross-classification of district and Rural/Urban type. Then, the sample within each stratum was allocated to the two sampling stages, i.e. number of schools to be selected, and number of students to be selected from each sampled school (or cluster of schools).

The number of students enrolled in Grade 3 obtained from the 2007/08 Census was the basis for allocating the sample across districts and Rural/Urban type. Since the sample had to be allocated simultaneously to the Province, Rural/Urban type within the province, and the 23 Districts in the province, a compromise allocation was used to allocate the total sample. This

⁵ The design effect is defined as the ratio of the variance under simple random sampling and the variance under the design that was actually implemented when the sample sizes are the same.

was aimed at striking a balance between conducting analyses simultaneously for the Province, districts and the rural/urban type.

The optimum number of students to be sampled per school was at most 10 students per school. The number of schools to be sampled from each primary stratum (i.e., Rural/Urban within each district) was computed by dividing the sample size in terms of number of students (obtained by raking procedure) by 10. The district with the smallest sample size was T.A. YAR with sample size of 103 schools. The sample size in terms of number of students was approximately 1,030 students in the smallest district.

The sample of the required number of schools was selected from each stratum with probabilities proportional to size (PPS), using the systematic sampling algorithm described in Hansen, Hurwitz, and Madow⁶ (1953). The measure of size (MOS) to be used for sample selection was the number of students in Grade 3 determined from the 2007/08 Census of Schools. The number of students in Grade 3 was used as the MOS because these students would be in Grade 4 at the time of testing and grade 4 was the target population.

It was important that the schools were sorted by Tehsil and Gender (Boys vs. Girls schools) within strata (the Rural and Urban parts of the Districts), and then by MOS by alternating between “ascending” and “descending” orders from one Gender type to the next. It should be noted that a school with 50 percent or more boys was defined as a Boy’s school, and the one with less than 50 percent boys was defined as Girl’s school.

As mentioned previously the schools with MOS (i.e. number of students in Grade 3 from the 2007/08 Census of Schools) less than or equal to 3 were not included in the survey population as it would not be very cost effective to sample very small schools. Ideally, the schools with MOS between 4 and 9 should have been collapsed. But, collapsing too many schools would result in operational issues, e.g. transporting students from several small schools to one test centre. As a compromise, the schools with MOS equal to 4, 5 or 6 were collapsed with neighboring schools within the same Union Council before sampling but the schools with MOS equal to 7, 8 or 9 were not collapsed. Thus, our primary sampling unit (PSU) was a cluster of schools instead of an individual school such that the PSU would have a minimum MOS of 7 students. The collapsing of the small schools would have been done using the criteria of minimum distance if Geographic Information System (GIS) was available. In the absence of GIS the small schools were collapsed within the Union Councils. A PSU that was a cluster of schools was treated as if it was a single school.

Conversely, if it happened that a school was so large that the corresponding selection probability became greater than one (selection probability must always be less than 1) it was decided to divide the original large school into a number of pseudo schools by a “conceptual split” where each pseudo school would be considered to be of the same size. The school was still one “physical” school and a 2nd stage sample of 10 students was selected from each sampled pseudo school. A “weight adjustment” was to be applied to account for the “conceptual

⁶ Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons

split” because the original school would now represent two or more pseudo schools. The weight adjustment factor was equal to inverse of the number of pseudo schools the “large” school was split into. If two or more pseudo schools got selected from the same “physical” school, then a separate sample of students was selected from the same “physical” school to represent each sampled pseudo school.

This sampling methodology of “conceptual split” was implemented so that the same survey processing system could be used for all schools including the “large” schools. It should be noted that a “large” school is not only large relative to other schools in the stratum but it also depends on the number of schools to be sampled from the stratum. Therefore, a school with certain MOS could be “large” in one stratum but a school in another stratum with even a greater MOS may not be “large” in that stratum.

The sample was designed using “EXCEL” and the sample of schools was selected using EXCEL as well.

The samples of students were selected by systematic sampling procedure by sorting the list of students in Grade 4 by section and by roll number within a section. Where the class list was less than or equal to 10 students, all students were selected. Otherwise, a sample of 10 students was selected with systematic sampling procedure. The systematic sampling procedure was implemented by providing the test administrators with random number tables with 10 sequence numbers out of the list of sequence numbers of the Grade IV students in the school. The random number tables were generated from 11 up to some maximum number of students (e.g. 400 students) in Grade 4 in the sampled school.

After the data collection and editing phases of the survey, the sampling weights for the data collected from the sampled students were constructed so that the responses could be properly expanded to represent the entire population of Grade 4 students in the government schools in the province of Sindh. The weights were the result of calculations involving several factors, including original selection probabilities, adjustment for non-response including both school non-response and student non-response, post-stratification adjustment based on the Grade 4 population of boys and girls within urban and rural parts in each district obtained from the 2008/09 Census of Schools.

Non-response is always present in any survey operation, even when participation is not voluntary. Thus, weight adjustment was necessary to account for the non-respondent schools and students. The non-response adjustment for the non-respondent schools was applied at the stratum level (calculated as the ratio of the MOS of schools (or clusters of schools) selected from the stratum and the MOS of those that participated in the assessment tests); whereas the non-response adjustment for the non-respondent students was applied at the school level (the weight adjustment was the ratio of number of sampled eligible students and the number that actually completed the assessment tests).

The base weight (or design weight) for each student was equal to the reciprocal of its probability of selection. The conditional selection probability of the student was equal to the number of students sampled divided by the number of students enrolled in Grade 4 in the school. An adjustment was made for the “large” schools that were split into pseudo schools. The adjustment factor to account for the “conceptual” split was equal to the number of pseudo schools that the large school was split into.

The post-stratification adjustment was applied by benchmarking the survey estimates for boys and girls within each primary stratum (District by rural/urban) to the enrollment obtained from the 2008/09 Census of schools.

The final survey weights for the respondent students were obtained as the product of the base weight, the two adjustment-factors for non-response (i.e. adjustment factor for non-respondent schools and adjustment factor for non-respondent students), and the post-stratification adjustment.

All survey estimates were obtained as domain estimates. The estimation domain was either a geographic domain (e.g., a district) or a characteristic domain (e.g., boys/girls). The estimation domain could also be the intersection of two or more geographic and/or characteristics domains, e.g. all boys in a particular district who achieved more than 80 percent scores. An indicator variable was used so that all estimates were expressed as “province” level estimates. The indicator variable automatically excluded those students that were not part of the estimation domain. The indicator variable technique ensures the proper estimation of variance.

Because the estimates were based on sample data, they differ from figures that would have been obtained from complete enumeration of the population of students using the same instrument. Results were subject to both sampling and non-sampling errors. Non-sampling errors included biases from inaccurate reporting, processing, and measurement, as well as errors from non-response and incomplete reporting. The non-sampling errors occurred at various phases of the survey process. However, to the extent possible, each error was minimized through the procedures used for data collection, editing, quality control, and non-response adjustment. The variances of the survey estimates were used to measure the sampling errors.

Quality assurance (QA) procedures were recommended and training was provided but these quality procedures were never implemented at all phases of the survey process. The main steps recommended for implementation for the QA procedure were: concept of a batch; verification of a sample of units from the batch; criteria to accept or reject the batch based on the observed error rate. Finally, the in-coming and out-going error rates were to be computed from the verification of the QA sample.

5.3 The Grade 4 Mathematics Provincial Sample

It was proposed that in 2008 a representative sample (15% of the Grade 4 population) of approximately 40,000 Grade four students (4333 PSUs) should participate in the assessment. Schools were selected, using Proportional Probability Sampling Techniques (PPS), in fixed proportions from the defined groups (districts; rural/urban; male/female).

Where schools in the sample had more than 30 Grade 4 students, the schools were split into PSUs; where the schools had less than 30 students, more than one school was collapsed to make one PSU.

The coverage of the provincial sample for Grade 4 Mathematics is found below. The actual sample used was less than the originally defined sample (28,684 students and 4333 PSUs). This was as a result of verification of the status of the schools by District Officials. The final sample used for analysis was 3,476 schools. This was a result of the data cleaning exercise conducted by PEACE.

The following table identifies the proposed number of schools to be sampled (Allocation from Sample Design), the sample schools selected, and the number of PSUs in the analysis. From this table it can be seen that the total number in the proposed sample is 4004 schools while the actual number of schools used in the analysis after data cleaning was 3476, a reduction of 528 schools.

Table 4: Sampled Schools and PSUs in Analysis

District_ID	District_Name	Location	Allocation from Sample Design	Sample Selected	PSUs in the Analysis	Diff	% Diff
1	Badin	Rural	124	120	110	14	11.3
		Urban	54	54	44	10	18.5
2	Dadu	Rural	144	144	136	8	5.6
		Urban	68	68	64	4	5.9
3	Hyderabad	Rural	30	28	26	4	13.3
		Urban	142	146	132	10	7.0
4	Thatta	Rural	144	144	127	17	11.8
		Urban	43	43	33	10	23.3
5	Mirpurkhas	Rural	92	92	81	11	12.0
		Urban	68	58	45	23	33.8
6	Tharparkar	Rural	135	133	90	45	33.3
		Urban	23	23	22	1	4.3
7	Sanghar	Rural	117	117	103	14	12.0
		Urban	84	85	79	5	6.0
8	Karachi	Rural	18	4	3	15	83.3
		Urban	305	318	289	16	5.2

12	Jacobabad	Rural	83	83	66	17	20.5
		Urban	58	56	49	9	15.5
13	Larkana	Rural	85	85	74	11	12.9
		Urban	117	117	113	4	3.4
14	Shikarpur	Rural	96	96	87	9	9.4
		Urban	62	62	61	1	1.6
15	Khairpur	Rural	174	172	153	21	12.1
		Urban	51	51	46	5	9.8
16	N_Feroze	Rural	153	153	141	12	7.8
		Urban	53	53	51	2	3.8
17	Nawabshah	Rural	110	109	87	23	20.9
		Urban	67	63	43	24	35.8
18	Sukkur	Rural	75	71	61	14	18.7
		Urban	93	99	97	-4	-4.3
19	Ghotki	Rural	170	170	129	41	24.1
		Urban	37	37	32	5	13.5
20	Umerkot	Rural	120	120	106	14	11.7
		Urban	20	19	19	1	5.0
22	Jamshoro	Rural	64	64	48	16	25.0
		Urban	56	56	31	25	44.6
23	Matiari	Rural	98	98	89	9	9.2
		Urban	34	33	32	2	5.9
24	T_Allahyar	Rural	79	79	75	4	5.1
		Urban	36	38	36	0	0.0
25	TM_Khan	Rural	71	71	62	9	12.7
		Urban	32	32	27	5	15.6
26	Kashmore	Rural	102	102	90	12	11.8
		Urban	41	41	35	6	14.6
27	Kambar	Rural	111	97	95	16	14.4
		Urban	65	65	57	8	12.3
Total			4004	3969	3476	528	13.2

The map below shows the total sample used in each district for assessment analysis.



5.4 Sampling Issues identified

The sample was selected using EXCEL. This did not automatically identify discrepancies such as, duplications of SEMIS Codes, replication of schools etc. During the identification of the sample no difficulties were identified. Few checks were conducted by the PEACE staff due to the lack of time available as the testing programme was required to be conducted within strict times with on-the-job training. Also PEACE did not have sufficient capacity or the available budget to enable rigorous monitoring and quality control.

Problems with the sample were identified through information provided by the Test Administrators and through using SAS for analysis.

This lack of implementation of quality procedures resulted in various types of errors being introduced:

- Some sample schools which showed 0 enrollment were found to have sufficient number of students for testing
- Some sample schools which showed high enrollment were found to have 0 enrollment

- Some schools which were required to be split because of large class enrolment, were not identified
- Approximately three schools which had been identified as requiring to be collapsed were found to be too large after collapsing and were then split (this should not have been done)
- Some school not in the identified sample were tested and the data received by PEACE
- Some schools which were in the sample did not send their data to PEACE
- Some schools which had been identified in 2007/08 Census having 0 enrollment and also schools which were identified as closed, were not part of the school sample despite having an appropriate student population

All of these errors resulted in the database being reduced.

5.5 Recommendations for Improvements in Sampling

The following are the recommendations for improvement:

- SAS software should be used for selecting the sample of schools as SAS code can be easily developed to identify and flag any discrepancies in the data.
- The technique developed during the sampling workshop to identify enrollment discrepancies, where the ratio of the student enrolment at the time of survey and the MOS is lower than 0.5 or higher than 1.5 should be used for further follow-up must be implemented. Where the ratio is found to be <0.5 or > 1.5 additional field checks will need to be made to ensure the reliability of the enrolment of the sampled schools.
- If the measure of size, which is the basis of the sample design, was larger than some threshold value, say 20, and it has changed greatly at the time of test administration it should be investigated. For example, if the MOS was 20 and it changes to 50, or if the MOS was 100 and changes to 20, these are serious discrepancies and should be investigated.

6. Large-scale Testing

Eight mathematics booklets A and B (two for each curriculum area) were finalised for large-scale testing covering the four curriculum content areas of number, fractions, measurement and geometry. Four subject specialists of PEACE, GEC and BoC finalised the instruments. The booklets contained 50 multiple choice items addressing different mathematical abilities. The item distribution by cognitive domains was as follows:

Table 5: Distribution of Test Items – Book A and Book B

Subject Area	Cognitive Domain		
	Conceptual Understanding	Procedural Knowledge	Problem Solving
Number Booklet A	16	24	10
Number Booklet B	20	25	5

Fractions Booklet A	15	25	10
Fractions Booklet B	15	25	10
Measurement Booklet A	22	18	10
Measurement Booklet B	13	27	10
Geometry Booklet A	26	18	6
Geometry Booklet B	21	24	5

6.1 District Focal Persons were identified in the District Education Department to be responsible for the efficient and smooth conduct of the testing and to be a bridge between the PEACE and the districts. A one day briefing was given to the focal persons regarding their roles and responsibilities. The main role of the focal persons was to organize the Test Administration training, distribution of assessment materials and the safe collection of the materials after the completion of the tests. The list of schools to be used in the tests was provided to the focal persons. At the successful end of the test administration activities the focal persons provided a certificate as evidence of the completion of all the testing tasks. A list of the District Focal Persons is in Annex 11.

6.2 The **printing** of the test booklets, background questionnaires, test administration guides, was undertaken by local printers in Hyderabad. Some difficulties occurred regarding the quality of the printing and where possible these were rectified. The test booklets are found in Annex 12; student, teacher and head teacher questionnaires in Annex 9 and test administration guidelines in Annex 10.

6.3 The **distribution** of assessment materials to the focal persons in the Districts, who were managing the further distribution of materials to the schools under tight deadlines, was a challenging task. All the materials arrived for the assessment but there were some delays and the delivery of duplicate materials as the deadlines for delivery were very narrow. A list of the Focal persons is found in Annex 11.

6.4 The administration of a test is a delicate procedure. To ensure the reliability and validity of the data in the field, uniformity of test administration is highly important. To ensure the standardisation of test administration, a one-day **training of Lead Master Trainers** was conducted. Ninety-six Lead Master Trainers were trained by PEACE at two different centres of Sindh, namely Sukkur and Hyderabad. A list of the Lead master Trainers is in Annex 13.

6.5 Training of Test Administrators was undertaken by the Lead Master Trainers and the training was monitored by PEACE and the district focal persons. The information regarding the numbers trained and the centres where this training took place and the number of test administrators trained is found in Annexures 14 and 15.

These master trainers trained test administrators throughout Sindh. The number of test administrators trained is found in the table below.

6.6 Some of the **difficulties** identified in the test administration were as follows:

- Some of the district focal persons and test administrators did not always appreciate the need for the assessment to be conducted in a rigorous manner;
- Test administrators did not successfully demonstrate the example questions in the test booklets to familiarize the students with the test methodology;
- Test administrators did not always follow the procedures given in the guidance booklet.
 - ✓ Test administrators found the use of the random number table (used to identify 10 students in a class of more than 10 students) as well as the skip interval difficult to understand and practise
 - ✓ There was also a lack of understanding of the methodology for entering the correct information for “split” schools
- The school enrollment was not recorded during the test administration for some of the sampled schools

7. Test Marking, Coding and Data Entry

For the 2009 mathematics test marking and coding methodologies were developed by the subject specialists on paper sheets and then transferred into the Excel program. Each possible answer was given a specific code. The markers did not mark questions right or wrong. If the first possible answer was chosen a code of 1 was given; for answer 2 a code of 2 was given; for answer 3 a code of 3 was given for answer 4 a code of 4 was given. Where a student had been given a misprinted test booklet or where the possible printed answers were not clear and the student has not answered the question, a code of 5 was allocated. Where a student marked two or more of the possible answers, a code of 7 was given; a code of 6 was given if the student was given a booklet of the incorrect language medium (Sindhi or Urdu). Where a student had not answered a code of 8 was given and where a student has not yet reached the question a code of 9 was given. This is in line with procedures developed in the National Education Assessment System.

Manual test marking and coding was conducted by elementary college, general school teachers and private school teachers at eight centres in the province (Annex 16). This involved marking and coding of approximately 800 assessment items for each student as well as 11 items in the Head Teacher’s Background Questionnaire, 54 items in the teacher questionnaire and 50 items in the students’ questionnaire. They were instructed on how to enter the data on the coding sheets before the start of the marking and coding process. They were paid for the completion of each booklet. The marking and coding was conducted in a timely manner. An example of the coding sheet used is found in Annex 17.

Checking the data was an onerous a task so it was not possible to check every single sheet The data of one student out of 10 students on a scoring sheet was checked by pairs of the elementary college, general school teachers and private school teachers and PEACE specialists super checked two out of 10 students on each scoring sheet in Hyderabad and the focal persons super checked two out of 10 students on each scoring sheet in the other centres. Where mistakes were found the students/teachers employed were asked to recheck their sheets and correct the mistakes. There appeared to be a lack of understanding of the need for

rigour in this work and it appears that the majority of the scorers and coders were mainly interested in the quantity of booklets they could complete rather than in doing the task well.

Twenty persons were involved in entering the data in Excel, 10 from the Bureau of Curriculum and 10 privately hired (Annex 18). They were instructed to enter data exactly as it was found on the scoring sheet. The data entered was checked in pairs by the data entry operators and super checking was conducted by three PEACE specialists and the remaining inconsistencies in the manual checking were identified and rectified. The Excel data was then converted to SAS software and SAS files were created.

Before converting the data to SAS software, the data was placed in four separate spread sheets for each of the 23 districts, namely, Booklet A rural, Booklet A urban; Booklet B rural, Booklet B urban. Recoding of the variable name and the removal of variables identified as not being useful for analysis purposes such as student starting and finishing times, enrollment, booklet serial number, book version etc. and the addition of a column for split schools, was undertaken. Fractional parts as found in the Excel files were re-coded from A, B, ...F to 1, 2, ...6. At the end of the conversion of A and B files, C and D files were obtained. Again many errors were found. Finally all district files were combined to make a whole province file for each of the four areas of mathematics tested and for each version (A and B). This resulted in eight data files. These C and D files were then compared to the A and B files of the sample frame files.

After this the SAS files were ready for the statistical analysis to be conducted.

At different stages of the process errors, inconsistencies and duplications were found.

The following mistakes were identified in the **manual entry** of information on to the score sheets:

- SEMIS Codes sometimes had digits missing or digits transposed or digits duplicated and split schools were often not entered accurately
- Coding errors were found regarding location (rural/urban)
- A few districts' names were not correctly entered
- Checking the gender code revealed some discrepancies
- Some schools' results were not found

Few mistakes were made in the **data entry** of the scores.

SAS software further identified the following issues:

- Duplication problems regarding the SEMIS Code and student roll numbers
- Split schools, fractional part was found to be incorrect and sometimes schools selected from the split school was without the fractional part and on occasion fractional parts were different from that in the sample
- The gender coding in two districts was found to be incorrect
- Discrepancies were found in the number of tests actually completed by students; these discrepancies ranged from one test to seven tests

- Schools did not match the sample

Marking and coding test booklets is an onerous task. Greater training needs to take place to ensure the validity of the information provided for analysis. The difficulties that have arisen from the methodology used will hopefully not occur in future assessments. SAS is now used and this program flags up any discrepancies immediately after data entry. The training of the markers and coders should be more thorough – besides explaining the methodology, trial runs of entering the data should take place and where the scorers and coders have difficulty their participation should be discontinued.

Background Questionnaires information was also inputted. The data was entered and cleaned on Excel and then converted to SAS. Some variables were found to be inconsistent and these were deleted. After converting to SAS a file was created for the whole province and data quality checked.

It was found that only nine questions out of the 54 questions in the student questionnaire were in usable form for analysis. The head teacher and teacher background questionnaires did not provide sufficient quality data for analysis and therefore were not used.

Some of the difficulties identified in the questionnaires were as follows:

- Some of the questions were translated incorrectly and provided little information
- There was a lack of specificity in some of the questions
- Many of the background questionnaires had non-response

For improvements in the response to the questions in the background questionnaires it is necessary for:

- Greater training to be imparted to Test Administrators
- More time to be provided for the completion of the questionnaires
- Review of the questionnaires to ensure improvements in the specificity of the questions
- Back translations to be conducted to ensure that the same questions are asked in the Sindhi and Urdu questionnaires

8. Data Cleaning

The purpose of data cleaning process was to ensure internal consistency of the data, and to check that the data would pass certain basic edits. The outcomes of data cleaning operations are reported in this section.

8.1 School Enrolment Data

A stratified sample of 3,969 PSUs was selected from 26,482 PSUs (File-A) with probability proportional to size (PPS) systematic sampling procedure. The 3,969 sampled PSUs correspond to 4,113 unique schools. The Grade IV enrolment in the sampled schools was used

to obtain the corresponding PSU level enrolment which in turn was used to compute the student sampling weights. The school enrolment data were reported by the Education District Officers (EDOs) for their respective sampled schools. If the enrolment reported by the EDO office was zero the school was automatically excluded from testing without any further verification. This is a biased procedure because the schools where the EDOs erroneously reported zero enrolment got excluded from testing and the corresponding PSUs were then categorized out-of-scope resulting in underestimation bias.

Table 6: Census 2008/09 vs. Survey Enrolment - Schools

School Enrolment		Number of Schools
Census 2008/09	Survey	
0	0	86
0	1	126
1	0	103
1	1	3,798
All Schools		4,113

Following this the ratio of the Survey enrolment reported by the EDOs and the Census 2008/09 enrolment for the **3,798** schools for which both enrolments (i.e. the Census 2008/09 enrolment and the survey enrolment) were non-zero, were computed. Seven classes were defined based on the ranges of the ratio of the survey and Census 2008/09 enrolments. The distribution of schools by class (i.e., different ranges of the ratio) is given in Table 7.

Table 7: Distribution of Schools by Ranges of the Ratio

Class	Range of Ratio	Number of Schools
1	Ratio \leq 0.50	426
2	0.50 < Ratio \leq 0.75	286
3	0.75 < Ratio \leq 0.95	282
4	0.95 < Ratio \leq 1.05	1,973
5	1.05 < Ratio \leq 1.25	205
6	1.25 < Ratio \leq 2.00	286
7	Ratio > 2.00	340
Total		3,798

Three hundred and forty schools were examined for which the survey enrolment was more than the double the enrolment reported by the Census 2008/09. Some of the ratios were so high that it was suspected that the enrolment figures reported by the survey were perhaps for the entire school and not just for the Grade IV class. Annex 19 provides the semis-codes of those schools for which the ratio was greater than 5.0 and the enrolment reported by the EDO office was more than 50. The ratios of the survey enrolment (reported by the EDO office) compared with that of the Census 2007/08 and with the Census 2008/09 are also given in Annex 19.

The PSU level enrolments were then computed for the 3,969 sampled PSUs from the school enrolment figures reported by the EDO and those obtained from the Census 2008/09, and checked whether the PSU enrolment computed from the Census 2008/09 and/or the one computed from the survey was zero or non-zero. The results of this check are provided in Table 8, where “0” implies zero enrolment and “1” implies non-zero enrolment.

Table 8: Census 2008/09 vs. Survey Enrolment - PSUs

PSU Enrolment		Number of PSUs
Census 2008/09	Survey	
0	0	56
1	0	77
0	1	90
1	1	3,746
All PSUs		3,969

Correlation with the MOS

The correlations between the PSU level MOS and the PSU level Survey enrolment, and the PSU level MOS and the PSU level Census 2008/09 enrolment using the PSU sampling weights were also computed. These correlations were respectively 0.41 and 0.85. It could not be possible that the correlation was so low between the MOS and the Grade IV enrolment that was reported by the survey because the MOS was the Grade III enrolment a year earlier. Therefore, these figures are for the same classes in two consecutive school years. On the other hand, the correlation between the MOS and the Census 2008/09 enrolment was 0.85 which is more realistic.

There were 90 PSUs for which Census 2008/09 enrolment was zero and the survey enrolment was non-zero, and tests had been administered in 60 out of these 90 PSUs. It would not have been possible to use the data for the students tested from these 60 PSUs if the Census 2008/09 enrolment had been used for sample weighting. To overcome this problem, the “modified

census” enrolment, for the purpose of sample weighting, was defined by taking the census enrolment if it was non-zero, and taking the survey enrolment if tests had been administered in the PSU. The “modified census” enrolment was set to zero for the remaining PSUs. The correlation between the MOS and the “modified census” enrolment was 0.86. Table 9 provides the comparison between the “modified census” enrolment, the Census 2008/09 enrolment and the survey enrolment for the 3,969 sampled PSUs.

Table 9: Modified Census, Census 2008/09 and Survey Enrolments - PSUs

PSU Enrolment			Number of PSUs
Modified Census	Census 2008/09	Survey	
0	0	0	56
0	0	1	30
1	1	0	77
1	0	1	60
1	1	1	3,746
All PSUs			3,969

Table 2-04 identifies that by using the “modified census” enrolment for the purpose of sample weighting only 86 PSUs were categorized out-of-scope as compared to 133 PSUs that would have been categorized out-of-scope if the enrolment reported by the EDOs had been used. The survey estimates would also have had additional bias because of erroneous enrolment figures reported by the EDOs. Moreover, the precision of the survey estimates would have been very low because of the poor correlation between the MOS and the survey enrolment figures provided by the EDO offices.

The files of 4,113 sampled schools were matched with that of the census 2008/09 because these counts were for the same Grade IV classes. Checking was undertaken to make sure that the reported school enrolment from the Census 2008/09 and/or the survey was zero or non-zero. The results of this check are provided in Table 2-01, where “0” implies zero enrolment and “1” implies non-zero enrolment

2.2 Student Test Results

There were 4 areas of assessment in mathematics: Fractions, Geometry, Measurement and Numbers. Two test versions (Book “A” and Book “B”) were administered for each of the 4 areas resulting in 8 different tests. Half of the students from a PSU would have attempted test versions “A” of the 4 areas and the remaining half would have attempted tests versions “B” of the 4 areas. Thus, a student could not have attempted both tests versions “A” and “B” for the same area of assessment. Moreover, a student could attempt a maximum of 4 tests only. The following two checks were applied to the student test data.

Students IDs with Both Tests A and B

Student IDs were matched to check if the same student had attempted both test versions “A” and “B” in the same area of assessment. It was found that there were 1,456 student IDs that did not pass this check. The reason for this discrepancy was that the same test results were data captured twice, once as Book “A” and once as Book “B”. The records for the incorrect version of the test were deleted. After data cleaning there were still 180 student records left that appeared in both Book “A” and Book “B” versions of the tests for the same areas of assessment. Because of time constraints these 360 records (180 records in each of the test versions “A” and “B”) were also deleted. Table 10 provides the distribution of students with both test versions “A” and “B” in the same area by area of assessment before and after data cleaning.

Table 10: Number of Students taking both Tests “A” and “B” in the same Area

Tests	Number of Students	
	Before Data Cleaning	After Data Cleaning
Both NA and NB	508	54
Both MA and MB	495	32
Both FA and FB	534	28
Both GA and GB	452	66
Total	1,456	180

Number of Students Tested and Number of Tests

There were 915 students on the data files with more than 4 tests. Some of these got reduced to less than or equal to four after keeping only one record for the students with both test versions “A” and “B” in the same area of assessment. There were additional students with more than 4 tests because the same student ID was assigned to two or more different students sampled from the PSUs formed by splitting a large school. There were also some records that were data captured multiple times. The data was cleaned by correcting the student IDs for the valid records or deleting the invalid records, e.g. duplicates. After data cleaning, 28,685 unique student data that attempted 106,716 tests was left. Thus, the average number of tests per student was 3.72 tests. Table 11 provides the distribution of students by number of tests taken.

Table 11: Distribution of Students by Number of Tests Taken

Number of Tests Taken	Number of Students
1	789
2	1,001
3	3,655
4	23,240
Total	28,685

As mentioned above, there were 8 different tests: FA (Fractions Book-A), FB (Fractions Book-B), GA (Geometry Book-A), GB (Geometry Book-B), MA (Measurement Book-A), MB (Measurement Book-B), NA (Numbers Book-A), and NB (Numbers Book-B). The distribution of the 106,716 tests by test is given in Table 12. The table provides the sample size in terms of the number of students for each of the 8 tests. These are the final sample sizes that were used for conducting data analyses.

Table 12: Sample Size (Number of Students) by Test

Test	Number of Students Tested
FA	13,388
FB	13,187
GA	14,616
GB	12,017
MA	13,847
MB	13,065
NA	13,299
NB	13,297
Total Number of Tests	106,716

From Table 12 it can be seen that there is a very good balance between the test versions A and B for Fractions, Measurement and Numbers. For the Geometry tests, the test version “A” has much larger sample than the test version “B”. It was found that only test version “A” was shipped to some of the schools and all students attempted test version “A” in those schools resulting in much larger sample size for the GA test as compared to the GB test. It should be noted that the sample sizes are also balanced across test areas, i.e. Fractions, Geometry, Measurement and Numbers.

Number of PSUs and Number of Tests

As noted above, a sample of 3,969 PSUs was selected, and a total of 28,685 students attempted 106,716 tests. Ideally, all 8 tests should have been administered in each eligible PSU but that was not case. Table 13 provides the distribution of PSUs by number of tests administered in these PSUs.

Table 13: Distribution of PSUs by number of tests

Number of Tests	Number of PSUs
0	493
1	29
2	32
3	36
4	179
5	176
6	295
7	631
8	2,098
Total	3,969

Table 13 identifies that there were 493 PSUs where no tests were administered. Eighty Six of these 493 PSUs were out-of-scope because there was zero enrolment in the Grade IV classes in these PSUs. Thus, there were 407 eligible PSUs where testing was not done. These 407 PSUs were categorized non-respondents for all 8 tests. Although tests were administered in 3,476 PSUs out of the 3,883 eligible PSUs, there were PSUs that were respondents for some tests but non-respondents for other tests because all 8 tests were not administered in these PSUs.

Number of PSUs at Allocation, Sampling and Data Analysis

The number of PSUs allocated to the strata at the design stage was compared with the number that was sampled and the number for which the data was available for conducting analyses. The distribution of PSUs by stratum at these 3 stages is given in Annex 20. The stratification was defined as cross-classification of district by location (Rural/Urban) resulting in 46 strata. It can be observed from Annex 20 that 13 percent of the PSUs between sample allocation and data analysis were lost. The most serious loss was for Karachi Rural where 83 percent of the PSUs were lost followed by Jamshoro Urban where 45 percent of the PSUs were lost. One third of the PSUs in each of the Mirpurkhas Urban, Tharparkar Rural and Nawabshah Urban strata were lost. These discrepancies are the result of lack of QA checks which had significant adverse impact on the quality of the data. In particular, the data for Karachi Rural has very low precision.

2.3 Background Questionnaire Data

There were three background questionnaires: Head Teacher, Teacher and Student. First, the three data files for missing IDs and duplicate IDs were checked. The records with missing IDs and duplicates were deleted. Table 14 gives the number of records on the original files and the number after removing the records with missing IDs and those with duplicate IDs.

Table 14: Number of Records on the Background Data Files

Background Questionnaire Data	Number of Records	
	Before Removing Missing/Duplicate IDs	After Removing Missing/Duplicate IDs
Head Teacher	2,666	2,560
Teacher	2,740	2,427
Student	27,217	26,890

The three files were linked to create a single student level file. The linked file had 27,183 student level records. The 27,183 records were then matched with the student file that contained student IDs of the students that took at least one test. It should be noted that there were 28,685 students with test data for at least one test. Out of the 27,183 background questionnaire records on the linked file only 22,016 records could be matched with the student file of 28,685 records. Thus, there were 6,669 student records with no background questionnaire data. The matched records had data from at least one of the background questionnaires, i.e. Head Teacher, Teacher or Student. The distribution of the 28,685 students by type of background questionnaire data is given in Table 15.

Table 15: Distribution of Students by Background Data Type

Background Questionnaire Data			Number of Students	
Student	Teacher	Head Teacher	Count	Percent
No	No	No	6,669	23.2
Yes	No	No	4,412	15.4
Yes	No	Yes	1,582	5.5
Yes	Yes	No	1,193	4.2
Yes	Yes	Yes	14,829	51.7
Total			28,685	100.0

The above table shows that there were only 14,829 out of the 28,685 (i.e., 51.7 %) student that had all 3 background data, i.e. Head Teacher, Teacher and Student. The above table only

provides the number of records with different types of background data. This does not imply that the data was valid as well. In fact most of the data was either missing or invalid. For example, it can be seen from the Table 2-10 that there were 22,016 student records (4,412 + 1,582 + 1,193 + 14,829) with the student background data. Nine data items (SQ10, SQ12, SQ15, SQ22, SQ24, SQ35, SQ38, SQ42 and SQ49) were checked for invalid and/or missing data values on the student background data file for those 22,016 students. There were only 12,106 student records out of the 22,106 matched records for which all 9 data items had valid responses. Thus, the student background data could only be used for 12,106 students if analyses had to be conducted with only the above 9 variables. The student background data for the 12,106 student were matched with the test results of these students.

It should be noted that the 12,106 matched students with the background questionnaire data had attempted multiple tests. The student background attributes were linked with all the tests that these students had attempted. We found that the 12,106 students had attempted 46,267 tests. Table 16 provides the distribution of the students with student background data by test (i.e., sample size for BQ data by test) along with the distribution of students that had attempted these tests (i.e., sample size for student test data by test) irrespective of whether the background data was available or not. It should be noted that the student background data were not available because either the student did not complete the background questionnaire or the student record could not be matched.

Table 16: Sample Size (Number of Students) by Test – Student BQ vs. Student Test

Test	Number of Students with BQ Data	Number of Students with Test Data	Ratio of Columns 2 and 3
Column 1	Column 2	Column 3	
FA	5,751	13,388	43.0
FB	5,782	13,187	43.8
GA	6,319	14,616	43.2
GB	5,153	12,017	42.9
MA	6,072	13,847	43.9
MB	5,631	13,065	43.1
NA	5,797	13,299	43.6
NB	5,762	13,297	43.3
Total Number of Records	46,267	106,716	43.4

Table 16 shows that only 43 percent of the students that attempted the tests had student background questionnaire data with valid data values. Moreover, the proportion of students with background questionnaire data was the same for all tests, i.e. 43 percent. Therefore, only 43 percent of the data was used to conduct analyses with the student background questionnaire data involving the 9 data items given above.

The data obtained from the Head Teacher and/or Teacher background questionnaires along with the student background questionnaire data had such a small number of records with valid data that no meaningful analysis could be conducted. For example, to use teacher background questionnaire data along with the student background data there were only 16,022 (i.e., 1,193 + 14,829) matched records, and the number of records with valid data depended on the data items used for the particular analysis. Therefore, analyses involving background questionnaire data would have been limited to very few data items.

The SAS code to match the background questionnaire data with the student test score data is given in Annex 21. The SAS code can also be used to check for valid/missing responses for different data items for further analyses.

Recommendations

- The grade IV enrolment for each sampled school must be verified by visiting the sampled schools to ensure that no in-scope school is excluded from testing. As the excluded schools get categorized as out-of-scope these results in an underestimation bias.
- The discrepancies in enrolment figures are the result of not implementing quality assurance (QA) checks during test administration. It is highly recommended that QA checks be implemented at each phase of the survey process.

9. Sample Weighting and Estimation

9.1 Sample Weighting

After creating the “clean” data files, weights were constructed for the students that participated in the assessment testing so that the responses could be properly expanded to represent the entire population of students that the sample was selected to represent. The sampling weights were the result of calculations involving several factors, including original selection probabilities, adjustment for non-response and benchmarking to the Census 2008/09 Grade IV enrolment as control totals. Since there were eight different tests eight sets of weights were created. The methodology for constructing the sampling weights was the same irrespective of the test.

PSU Weight

The base weight (or design weight) for a sampled PSU is the reciprocal of the PSU selection probability. We denote by w_{hi} the base weight of the PSU i selected from the stratum h . It should be noted that the stratification was defined as cross-classification of district by location (Rural/Urban) resulting in 46 design strata. The sampled PSUs were then divided into 3 categories:

1: Respondents. This category consists of eligible sampled PSUs (schools) that participated in the testing and provided usable survey responses. This category is denoted by R.

2: Non-respondents. These are the eligible sampled PSUs (schools) that did not participate in the testing for the particular area of assessment. This category is denoted by N.

3: Out-of-Scope. These are the sampled PSUs (schools) that were not eligible for the survey because either there were no Grade IV students in the school or the school did not exist. This category is denoted by O.

We also denote by x_{hi} the measure of size (MOS) of the sampled PSU i from the stratum h . Then, the non-response adjustment factor at the stratum level was computed as:

$$(Adj_Non_Response)_h = \frac{\sum_{hi \in R} w_{hi} x_{hi} + \sum_{hi \in N} w_{hi} x_{hi}}{\sum_{hi \in R} w_{hi} x_{hi}}. \quad (3-1)$$

The non-response adjusted PSU weights were then computed as the product of the PSU base weights and the corresponding non-response adjustment factors.

Student Weight

The conditional probability of selecting a student given that the PSU had been sampled is defined as the ratio of the number of students sampled and the number enrolled in Grade IV in the PSU. The student weight is the reciprocal of the conditional probability of selection. Since all students may not participate in the test, non-response adjusted student weight was defined as the ratio of the number of students enrolled in the PSU and the number that actually participated in the test for the particular area of assessment. It should be noted that where two or more schools were collapsed to form a PSU the PSU enrolment was the sum of enrolments of all schools that were collapsed to form the PSU. On the other hand, in the case of splitting a large school the PSU enrolment was adjusted by dividing the school enrolment by the number of splits.

Overall Student Weight

The overall (or initial) student weight was computed as the product of the non-response-adjusted PSU weight and the non-response-adjusted conditional student weight. It should be noted that the initial student weight will be the same for all students tested from the same PSU.

Post-stratification Adjustment

Post-stratification is a popular estimation procedure in which the initial weights after non-response adjustment are further adjusted so that the estimated totals based on the final weights (i.e. after post-stratification adjustment) are equal to known population totals (or more precise estimates of the population totals) for certain subgroups of the population. Since student enrolment figures were available from the 2008/09 annual census of schools we used the Grade IV census enrolment as control totals for post-stratification.

The post-strata was defined as the cross classification of district by location (Rural/Urban) by gender (Boys/Girls) resulting in 92 post-strata. Let Z be the Grade IV enrolment obtained from the 2008/09 annual census of schools. We denote by Z_g the total of the variable Z for the post-stratum g ($g = 1, 2, 3, \dots, 92$) that is based on the 2008/09 annual census of schools. Similarly, we denote by \hat{Z}_g the estimated total that is based on the overall (initial) student weights. The post-stratification adjustment is then defined as the ratio Z_g / \hat{Z}_g .

Annex 222 provides information on post stratification control totals (Enrolment).

Calculation of the Final Student Weights

The post-stratified weights for the students that were administered a particular test, e.g. FA, were calculated as the product of the overall (initial) student weights and the corresponding post-stratification adjustment factors. The post-stratified weights were the final weights of the students that were used to construct estimates for various estimation domains and to conduct tests of hypotheses.

Because a student can belong to one and only one of the post-strata, the post-stratified weights are uniquely defined. The advantage of post-stratified weighting is that the reliability of the survey estimates is improved when there is high correlation between the auxiliary variable used for post-stratification and the study variable. Moreover, most of the bias due to survey under-coverage gets corrected.

It had been planned to apply gender factors at the stratum level to correct for any deviations between the estimated boys and girls counts from the sampled PSUs and the numbers known from the frame (population). The gender factors became redundant when post-stratification adjustment factors were applied, and gender is incorporated in the definition of post-strata.

9.2 Survey Estimates

All survey estimates were obtained as domain estimates by using an indicator variable ${}_d\delta_i$, where the post-script d denotes the “estimation domain” and the sub-script i denotes the respondent student. It should be noted that the sub-script i now denotes a unique student within the province. The estimation domain can be a geographic domain (e.g., a district) or it can be a characteristic domain (e.g., boys/girls). The estimation domain can also be the intersection of two or more geographic and/or characteristics domains. For example, all girls in a particular district who obtained 80 percent or higher score. The indicator variable ${}_d\delta_i$ is defined as:

$${}_d\delta_i = \begin{cases} 1, & i \in d \\ 0, & \text{Otherwise} \end{cases} \quad (3-2)$$

Then the estimated total of the study variable Y for the domain of interest d can be expressed as:

$${}_d\hat{Y} = \sum_{i \in S} w_i^* \times {}_d\delta_i \times y_i, \quad (3-3)$$

where y_i is the reported value (e.g. student score) of student i , and w_i^* is the corresponding final student weight. The summation symbol $\sum_{i \in S}$ denotes the summation over all selected students that provided useable data. The indicator variable ${}_d\delta_i$ defined in equation (3-2) would include the contribution only from those students that belong to the estimation domain.

The advantage of using the indicator variable is that all estimates can be expressed as the “province” level estimates. The indicator variable ${}_d\delta_i$ will automatically exclude those students that are not part of the estimation domain.

9.3 Quality of the Survey Estimates

Because estimates are based on sample data, they will differ from figures that would have been obtained from testing the entire population of Grade IV students using the same instrument. Results are subject to both non-sampling and sampling errors. Non-sampling errors include biases from inaccurate reporting, processing, and measurement, as well as errors from non-response and incomplete reporting. These types of errors cannot be measured readily. However, to the extent possible, each error can be minimized by implementing quality control procedures at various phases of the survey operation. The variances of the survey estimates are used to measure the sampling errors. The variance estimation methodology is discussed in this section.

Variance Estimation

Use of standard statistical techniques is not appropriate for analyzing data collected in complex surveys. Estimates from complex survey data, like ratio means, odds ratios, regression coefficients, etc. are themselves complicated, and methods of standard error estimation are needed that account for these complexities. The software WesVar⁷ was used for variance estimation, and testing hypotheses. WesVar computes estimates and replication variance estimates that do properly reflect complex sampling and estimation procedures.

Fay’s replication method in WesVar was replicated because of its simplicity and its properties for estimating variances of non-linear statistics. The Fay’s method is applicable when two primary sampling units (PSUs) are sampled from each stratum. The selection of PSUs with PPS systematic sampling procedure by using a sort ordering that provided implicit stratification. Therefore, consecutive groupings of sampled PSUs were treated as pseudo stratum.

⁷ WesVar is Statistical Software developed by Westat, Inc. for analyzing data from complex surveys.

Pseudo strata were created by taking groups of 16 consecutive PSUs. The pseudo strata were created within design strata. Therefore, the last group did not always consist of 16 PSUs. Each group can be treated as a pseudo stratum because of implicit stratification by sort ordering for the PPS systematic sampling. The two variance units within pseudo strata were created by using even vs. odd PSUs.

246 pseudo strata (VarStrat) were created with two variance units (VarUnits) for implementing Fay's replication method in WesVar. 256 replicate weights were created and the same post-stratification adjustment was applied to each of the replicate weights as the full sample weights. Thus, the non-linear adjustment was reflected in the variance estimation.

Other Measures of Precision

In practice, the sampling variance is hardly ever reported. Instead, users find it more useful to rely on one of the derivatives of the sampling variance, such as the *standard error*, the *coefficient of variation*, the *margin of error*, or the *confidence interval*. These are all related expressions, and it is quite easy to go from one to the other using simple mathematical operations.

A. Standard Error

The standard error of an estimator is the square root of its sampling variance. This measure is easier to interpret since it provides an indication of sampling error using the same scale as the estimate whereas the variance is based on squared differences.

If $\hat{\theta}$ is the estimate of an arbitrary population parameter θ and $v(\hat{\theta})$ is the corresponding estimate of its variance, then the standard of the estimate is defined as:

$$se(\hat{\theta}) = \sqrt{v(\hat{\theta})}. \quad (3-5)$$

B. Coefficient of Variation

It is more useful in many situations to assess the size of the standard error relative to the magnitude of the characteristic being measured. The **coefficient of variation** (cv) provides such a measure. It is the **ratio of the standard error of the survey estimate to the value of the estimate itself expressed as percent**. It is very useful in comparing the precision of several different survey estimates, where their sizes or scale differ from one another. The coefficient of variation of $\hat{\theta}$ denoted by $cv(\hat{\theta})$ is defined as:

$$cv(\hat{\theta}) = 100 \times \left\{ \frac{se(\hat{\theta})}{\hat{\theta}} \right\}. \quad (3-6)$$

C. Construction of Confidence Intervals

The 95 percent confidence interval is the interval such that there is a 95 percent probability (chance of 19 out of 20) of the unknown population parameter θ to be within the interval.

D. Design Effects

Most surveys are based on complex designs involving stratification, and clustering due to multi-stage designs. Moreover, the weighting involves non-linear adjustments (e.g., non-response and post-stratification adjustments, etc.). It is crucial that these features of the complex survey design be accounted for in the variance estimation. The **design effect** compares the variance of the estimate from the sample design that was actually implemented to the variance of the estimate that would have been obtained from an SRS design. It is defined as the ratio of the variance of an estimate for a complex sample design and the variance of the estimate under the simple random sample (SRS) design with the same sample size. **Design Effect** is another way to evaluate the efficiency of a sample design and the procedure used to develop the survey estimates. It is important to note that the design effect is associated with both the design and the estimator; therefore, for a given survey, the design effect can vary quite a lot from one variable to another.

E. Effective Sample Size

Another concept that is often used is **effective sample size** defined as the actual sample size that was selected for the complex design divided by the corresponding design effect. The effective sample size can be interpreted as the sample size that would be needed for an SRS design to obtain the same variance as that obtained with the complex design (i.e. the design that was actually implemented).

10. Data Analysis

Data analyses were conducted both at the district level and the province level. The tests of hypotheses were conducted using t-tests to test the significance of differences between two estimates. There are 3 alternate approaches that can be used for doing a t-test. These are equivalent approaches and the resulting inference is the same. These are described below.

1. **95 % Confidence Interval:** If the value “zero” lies inside the 95% confidence interval of the estimated difference (i.e., lower limit is negative and the upper limit is positive) then the observed difference is not significant at the 95% confidence level. If both the lower and the upper limits are negative, or these limits are both positive then the difference is significant in the corresponding direction.
2. **T-Vale:** We recall that the variances of the survey estimates were computed with Fay’s replication method, and there were 256 replicates. The t-value at $\alpha=0.05$ with 256 degrees of freedom is 1.969. Therefore, if the observed t-value lies between - 1.969 and +1.969 then the estimated difference is not significant. On the other hand, if the observed t-value is less than -1.969 or greater than +1.969 then the estimated difference is significant in the corresponding direction.
3. **P-Value:** The p-value is the probability that a t-value more extreme than the one actually observed can occur under the null hypothesis of no significant difference. If the probability is greater than 0.05 then the observed difference is not significant. On the other hand, if the probability is less than or equal to 0.05 then the observed difference is significant. If the probability is less than or equal to 0.01 then the observed difference is also referred to as highly significant.

Chi-square tests were also used for testing independence in two-way tables, e.g. test versions “A” and “B” by ability level where ability level was defined with 4 categories of ability. The Rao-Scott⁸ chi-square was used because the chi-square accounts for the complex sample design. Below are the results of these data analyses.

10.1 District Level Analyses

As mentioned before, there were four areas of assessment and two test versions were administered for each area resulting in 8 tests. The differences between the average percent scores for boys and girls for each of the 8 tests at the district level were computed and t-tests for significance of differences between achievements (i.e. average percent scores) of boys and girls were conducted. The results are given in Tables A1.1 to A1.8 in Annex 23. The differences between the achievements of boys and girls for the aggregate of all 8 tests at the district level

⁸ Rao, J.N.K., and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association*, **76**, 221-230.

were computed and t-tests conducted to test for significance of their differences. The results are given in table A1.9 (Annex 23). Similarly, the differences between the average percent scores of the students in the rural schools and those in the urban schools for the 8 tests and the aggregate of all 8 tests by district, and the results of the corresponding t-tests are given in tables A2.1 to A2.9 (Annex 23). From these tables it can be observed that the differences between boys and girls average percent scores were not significant for most districts. The ones where the differences were significant these were positive for some districts and negative for others. For example, the differences between boys and girls for the aggregate of all 8 tests were significantly positive for Sukkur, Ghotki and Jamshoro, and these differences were significantly negative for T.M. Khan and Kashmore. It should be noted that positive difference implies that the boys' average percent score was higher than that of girls and vice versa. Similarly, the differences between rural and urban students for the aggregate of all 8 tests were significant positive for Jacobabad, Sukkur, Ghotki, Umerkot and Kashmore, and the difference was significantly negative for Kambar.

Table A-3.0 in Annex 8 provides the average percent score (column 2) for the aggregate of all 8 tests by district, and percent boys (column 4) and percent rural students (column 5) in each district. It can be observed that some of the districts with above average achievements have high percent of boys (e.g. Kashmore, Ghotki, T.M. Khan, Mirpurkhas, Sanghar and Jacobabad), and some others with below average achievements have low percent of boys (e.g. Hyderabad and Karachi). The net effect is that the average percent score for boys at the province level is significantly greater than that of girls. It is the same phenomenon for the rural/urban students. Most of the districts with above average achievements have high percent of rural students (e.g. Kashmore, Ghotki, Tharparkar, N. Feroze, T.M. Khan, Mirpurkhas, Sanghar, Jacobabad, Matiari and Dadu), and some others with below average achievements have low percent of rural students (e.g. Hyderabad, Karachi, Sukkur and Larkana). The net effect is that the average percent score for rural students at the province level is significantly greater than that of the urban students. Thus, we can conclude that the real differences in achievement are across districts, and the observed significant differences between boys and girls and rural and urban students are mere manifestations of the differences across districts.

Districts vs. Rest of the Province

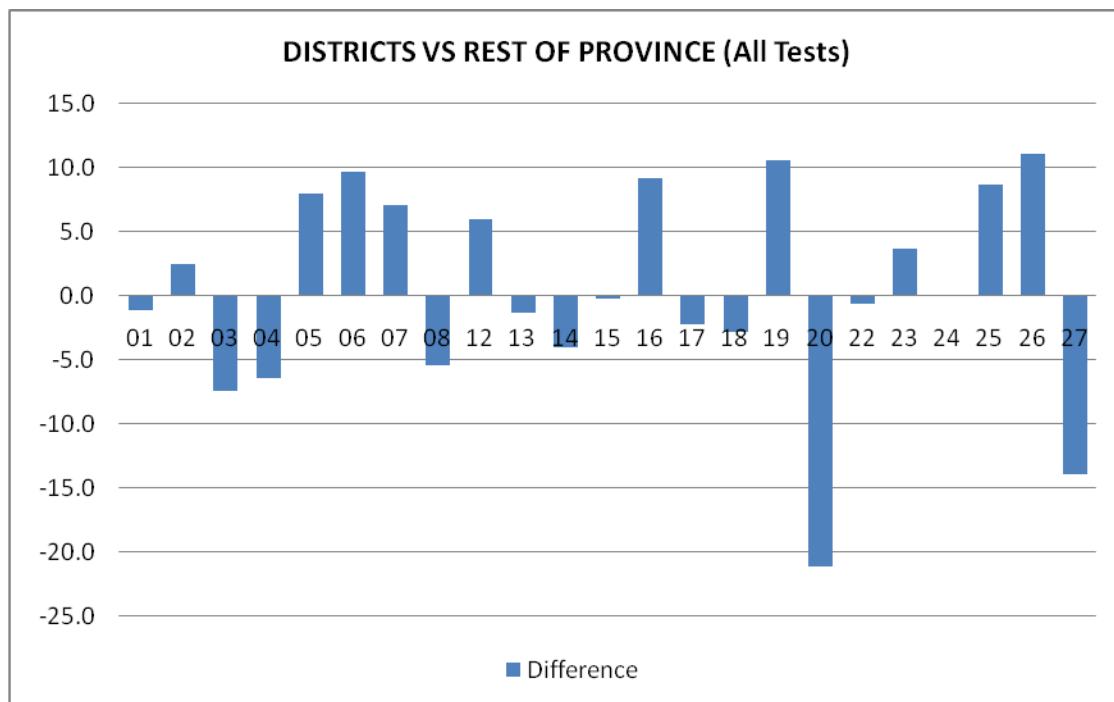
District level achievements were compared with the rest of the province by computing the differences between the average percent scores for the districts and for the rest of the province for the aggregate of all 8 tests. T-tests were conducted to test the significance of these differences. The results are reported in Table A-4.0 in Annex 23.

Table A-4.0 shows that the overall average achievements of the districts of **Kashmore, Ghotki, Tharparkar, N. Feroze, T.M. Khan, Mirpurkhas, Sanghar, Jacobabad and Matiari** were significantly higher than the corresponding rest of the province averages. On the other hand, the overall average achievements of the districts of **Umerkot, Kambar, Hyderabad, Thatta, Karachi, Shikarpur and Sukkur** were significantly lower than the corresponding rest of the province averages. There was no significant difference between the average district level

achievements and the corresponding rest of the province average for the other districts (i.e. **Nawabshah, Larkana, Badin, Jamshoro, Khairpur, T. Allahyar and Dadu**).

The graph below shows the differences between the average percent scores of the districts and that of the rest of the province for the aggregate of all 8 tests. A positive difference implies that the average percent score of the district was higher than that of the rest of the province, and a negative difference implies that the average percent score of the district was lower than that of the rest of the province. The districts of **Umerkot and Kambar** are the two districts with the lowest average percent scores, and the districts of **Kashmore and Ghotki** are the two districts with the highest average percent scores.

Difference between Districts and Rest of Province Average percent Scores (All Tests)



Legend:							
01=Badin	02=Dadu	03=Hyderabad	04=Thatta	05=Mirpurkhas	06=Tharparkar	07=Sanghar	08=Karachi
12=Jacobabad	13=Larkana	14=Shikarpur	15=Khairpur	16=N. Feroze	17=Nawabshah	18=Sukkur	19=Ghotki
20=Umerkot	22=Jamshoro	23=Matiari	24=T. Allahyar	25=T.M. Khan	26=Kashmore	27=Kambar	

Chi-Square test

In order to conduct chi-square tests of independence for two-way tables, four ability levels were defined based on the percent scores as follows.

Ability Level 1: Percent Score ≥ 80.0 %;

Ability Level 2: 68.0 % \leq Percent Score < 80.0 %;

Ability Level 3: 40.0 % \leq Percent Score < 68.0 %; and

Ability Level 4: Percent Score < 40.0 %.

A chi-square test of independence in the two-way table of gender (Boys/Girls) by ability level for the aggregate of all 8 tests was conducted. The results are given in Table 17. Similarly, a chi-square test of independence in the two-way table of location (Rural/Urban) by ability level for the aggregate of all 8 tests was conducted, for which the results are given in Table 18.

Table 17: Chi-Square Test for Independence of Distribution of Ability Level by Gender

Test	Gender	Ability Level				Chi-Square	DF	p-Value
		Level1	Level2	Level3	Level4			
All Tests	Boys	10.5	15.4	16.7	57.4	29.3	3	0.00
	Girls	9.1	13.5	16.0	61.4			

Table 18: Chi-Square Test for Independence of Distribution of Ability Level by Location

Test	Location	Ability Level				Chi-Square	DF	p-Value
		Level1	Level2	Level3	Level4			
All Tests	Rural	10.8	15.4	16.4	57.4	49.7	3	0.00
	Urban	8.0	12.8	16.5	62.7			

The chi-square values are highly significant for both tests which are consistent with the earlier findings from the t-tests. Even though the chi-square values are highly significant, the significant differences between the distributions by ability levels of boys and girls, and those between the distribution by ability levels of rural and urban students are actually caused by differences in achievements across districts.

10.2 Province Level Analyses

t-tests were conducted to compare test versions “A” and “B” for each of the four areas of assessment. Moreover, pair-wise comparisons (t-tests) among the four areas of assessment were also done. t-tests to compare the average percent scores of students for small vs. large class size were also conducted. Estimates of average percent scores of “Contextual” and “Non-contextual” test items by gender and by location were computed. The items were also categorized into 3 categories related to “Conceptual Understanding”, “Problem Knowledge” and “Problem Solving”. The estimates of average percent scores of the above 3 categories were computed by gender (Boys/Girls) and by location (Rural/Urban).

Test versions A and B

Test versions A and B were compared for each of the four areas of assessment by gender (Boys/Girls) and by location (Rural/Urban). The results for the Boys and Girls are given in Table A-5.1 (Annex 23) and those for Rural and Urban students are given in Table A-5.2 (Annex 23).

From Table A-5.1 it can be observed that there is no significant difference between the achievements of students attempting test version “A” and those attempting test version “B” of Fractions both for Boys and for Girls. The differences between test versions “A” and “B” of

Geometry, Measurement and Numbers are highly significant both for Boys and for Girls. Achievements both for Boys and for Girls are better with test version "A" than with test version "B" for Measurement. Achievements both for Boys and for Girls are better with test version "B" than with test version "A" for the other two areas of assessment (i.e., Geometry and Numbers).

It can also be observed from Table A-5.2 that there is no significant difference between the achievements of students attempting test version "A" and those attempting test version "B" of Fractions for the students in the rural schools. All other differences between test versions "A" and "B" are highly significant. The students in the urban schools had better achievement with test version "A" than with test version "B" of Fractions. Achievements both for rural students and for urban students were better with test version "A" than with test version "B" of Measurement. Achievements both for rural students and for urban students were better with test version "B" than with test version "A" of the other two areas of assessment (i.e., Geometry and Numbers).

Chi-square tests of independence in the two-way tables of test version (A vs. B) by Ability Level for the four areas of assessment were conducted. The results are given in Table 4-3.

From Table 19 it can be seen that the chi-square is not significant for Geometry, but the chi-square values are highly significant for the other 3 areas of assessment (i.e. Geometry, Measurement and Numbers). This result is consistent with the results of the t-tests in Tables A-5.1 and A-5.2 in Annex 23.

Table 19: Chi-Square Test for Independence of Distribution of Ability Level by Tests A & B

Area	Test	Ability Level				Chi-Square	DF	p-Value
		Level1	Level2	Level3	Level4			
Fraction	A	9.1	9.6	11.3	70.0	7.41	3	0.06
	B	7.6	10.4	11.4	70.6			
Geometry	A	3.9	11.3	19.6	65.2	263.68	3	0.00
	B	4.6	21.0	25.7	48.7			
Measurement	A	12.9	19.4	17.9	49.8	89.70	3	0.00
	B	11.6	12.9	17.9	57.6			
Numbers	A	16.0	13.4	12.1	58.5	97.25	3	0.00
	B	13.7	19.0	15.5	51.8			

Pair-wise comparison of the four areas of assessment

t-tests to do the pair-wise comparisons among the four areas of assessment by gender and by location were conducted. The results by gender are given in Table A-6.1, and those by location are given in Table A-6.2 (Annex 23). The achievement in Fractions was significantly lower than that in all other areas of assessment (Geometry, Measurement and Numbers) for Boys, Girls, rural students and urban students. As a result, the achievement in Fractions was also significantly lower than that in the other 3 areas at the province level. Also, the achievement in Geometry was

significantly lower than that in Measurement and Numbers for Boys, Girls, rural students and urban students, and hence at the province level. The differences between Measurement and Numbers were not significant for any of the subgroups or at the province level.

Small vs. Large Class Size

Two categories of class size were defined. The class size was defined as “small” if the number of students in the class was less than 15, and it was “large” otherwise. t-tests to test the significance of differences in achievements of small vs. large class sizes for boys and girls, and for all students were conducted. The results are given in Table 20. It was found that there was no significant difference between the achievements of students in small vs. large class sizes for boys, girls and the total. Similarly, the results for rural and urban students are given in Table 21 it can be observed that there was no significant difference between the achievements of students in small vs. large size classes for rural school students and urban school students.

Table 20: Test of Significance of Difference between Small and Large Class Sizes (Boys and Girls)

Gender	Small Size		Large Size		Difference Columns 2 & 4	95% Confidence Interval of Col. 6		T Value	Probability
	Estimate	% CV	Estimate	% CV		Lower	Upper		
1	2	3	4	5	6	7	8	9	10
Boys	45.7	1.2	45.5	1.1	0.2	-1.5	1.8	0.20	0.85
Girls	44.6	1.7	43.0	1.3	1.7	-0.3	3.7	1.65	0.10
SINDH	45.3	1.1	44.5	1.0	0.8	-0.6	2.2	1.18	0.24

Table 21: Test of Significance of Difference between Small and Large Class Sizes (Rural and Urban)

Location	Small Size		Large Size		Difference Columns 2 & 4	95% Confidence Interval of Col. 6		T Value	Probability
	Estimate	% CV	Estimate	% CV		Lower	Upper		
1	2	3	4	5	6	7	8	9	10
Rural	45.5	1.2	45.4	1.6	0.1	-1.5	1.6	0.05	0.96
Urban	43.4	2.6	43.1	1.1	0.3	-2.2	2.9	0.25	0.80
SINDH	45.3	1.1	44.5	1.0	0.8	-0.6	2.2	1.18	0.24

Contextual vs. Non-contextual Items

The items in each test were divided into two categories: Contextual and Non-contextual test items. The contextual items are those for which students had to figure out the mathematical operation to be applied to obtain the answer. Whereas, the non-contextual items are the items where students were asked to apply a given mathematical operation. Estimates of average percent scores for Contextual vs. Non-contextual test items in each of the four areas of assessment and their aggregate for boys and girls were obtained. The estimates are given in Table A-7.1 in Annex 23. It can be observed that the achievement is very low for the contextual items as compared with that for the non-contextual items in each area of assessment and aggregate of the areas both for boys and for girls, and hence for both collectively.

Similarly, estimates of average percent scores for Contextual vs. Non-contextual items in each of the four areas of assessment and their aggregate for rural students and for urban students were obtained. The estimates are given in Table A-7.2 in Annex 23. It can be observed that the achievement is very low for the contextual items as compared with that for the non-contextual items in each area of assessment and the aggregate of the 4 areas both for rural students and urban students.

Conceptual-understanding, Problem knowledge and problem solving

The items in each test were divided into 3 categories related to: Conceptual Understanding, Problem Knowledge and Problem Solving. The estimates of average percent scores for the above 3 categories in each of the four areas of assessment and their aggregate for boys and for girls are given in Table A-8.2 in Annex 23. Similarly, the average percent scores for the 3 categories in each of the four areas of assessment and their aggregate for rural students and for urban students are given in Table A-8.2 in Annex 23. It can be observed that the achievement is the highest in the test items related to Conceptual Understanding in all four areas of assessment for all groups, e.g. boys, girls, rural students, urban students. The achievement was higher in the test items related to Procedural Knowledge than those related to Problem Solving for Geometry, Measurement and Numbers. But, the achievement was slightly lower for test items related to Procedural Knowledge than those related to Problem Solving in Fractions.

10.3 Analyses with Background Questionnaire Data

First, it should be noted that we have background questionnaire data for only about 43 percent of the students out of those that attempted the assessment tests. Therefore, the sampling weights (full sample weights as well as the replicate weights) had to be reconstructed so that the sample would represent the entire population of grade IV students. Moreover, pseudo strata (VarStrat) and variance units (VarUnits) for implementing Fay's replication method had to be defined again because of the reduced sample size. The number of pseudo strata reduced from 246 to 101, and a set of 104 replicate weights were constructed using Fay's replication method.

The background questionnaire data for 9 student items (SQ10, SQ12, SQ15, Sq22, SQ24, SQ35, SQ38, SQ42 and SQ49) was used for conducting analyses. It should be pointed out that there would be the risk of potential bias if the students that provided the background data differed systematically from those that did not complete the background questionnaires or those that could not be matched with the background questionnaire data file.

The nine student background questionnaire data items were taken from three different groups of questions on the student background questionnaire. These are described below.

Group 1

HOW OFTEN DO YOU SPEND TIME WITH A PARENT OR OTHER ADULT FAMILY MEMBER OUTSIDE SCHOOL IN THE FOLLOWING WAYS?

- SQ10: Talking about books or studies.
- SQ12: Playing sports or games or keeping fit.

Answer Categories:

1=Daily; 2=Weekly; 3=Once a month; 4=Two or three times during the year; 5=Seldom.

Group 2

WHAT ARE MATHEMATICS LESSONS LIKE?

- SQ15: Answers to our questions are explained in detail.
- SQ22: Most of the assessment in mathematics is done on short tests.
- SQ24: We get regular homework in mathematics.
- SQ35: My family thinks that mathematics is an important subject.

Answer Categories:

1=Yes, Always; 2=Mostly; 3=Not usually; 4=Definitely Not; 5=Do not know.

Group 3

DURING MATHEMATICS LESSONS HOW OFTEN DO YOU SPEND YOUR TIME ...?

- SQ38: Being taught by your teacher in small groups.
- SQ42: Working quietly on your own.
- SQ49: Using real-life examples to work with.

Answer Categories:

1=Most lessons; 2=Most weeks; 3=Once or twice each term; 4=Once a year or less.

Analyses were conducted by collapsing the last two answer categories for each of the 9 data items. Thus, the recoded SQ10, SQ12, SQ15, SQ22, SQ24 and SQ35 would have 4 categories, and the remaining 3 data items (SQ38, SQ42 and SQ49) would have 3 categories.

Chi-Square Test of Independence

For each of the 9 items we performed chi-square test of independence in a two-way table with the ability level. The results are provided in Table A-10.0 in Annex 23. It should be recalled that the categorical variable ability level has 4 categories. Thus, the two-way table of SQ10 by Ability Level is a 4 x 4 table, and under the null hypothesis the test statistics will be distributed as chi-square with 9 degrees of freedom. It can be observed from Table A-10.0 that chi-square is not

significant for SQ12 and SQ42. It is significant for SQ10 and SQ49; and highly significant for SQ15, SQ22, SQ24, SQ35 and SQ38. For example, ability level is higher for SQ38 = 2 (i.e., taught in small groups during most weeks) as compared to the other two categories, i.e. SQ38 = 1 (i.e., taught in small groups during most lessons) and SQ38 = 3 (i.e., seldom taught in small groups).

Linear Regression Model

An estimated linear regression model using the percent score as the dependent variable, and the categorical variables District-ID, Gender, Location, SQ15 and SQ38 as explanatory variables was used. The intercept term in the model was used. As expected, the intercept term was positive and highly significant. The results for the other variables were as follows.

District-ID

There are 23 districts in the province and the district with district-ID = 27 (Kambar) was the reference category. The coefficient was negative for the district-ID = 20 (Umerkot) and the coefficients for all remaining 21 were positive but all 22 coefficients were highly significant. This is consistent with the earlier finding that Umerkot has the lowest achievement followed by Kambar, and all other districts have higher achievements than these two districts.

Gender

The category Gender = 2 (Girls) was the reference category and the coefficient for the category Gender = 1 (Boys) was not significant (p-value was 0.12). This shows that the real differences in the achievement are across districts, and there is no significant difference between boys and girls after the categorical variable "District" has been included in the model.

Location

The category Location = 2 (Urban) was the reference category and the coefficient for the category Location = 1 (Rural) was not significant (p-value was 0.91). This shows that there was no significant difference between Rural and Urban students either after the categorical variable "District" has been included in the model.

SQ15-Recoded

The SQ15 question was: Answers to our questions are explained in detail. The category SQ15-Recoded = 4 (Never or Do not know) was the reference category. The coefficients for all other categories were positive and significant. The values of the three coefficients, and t-values and p-values for testing the null hypothesis that coefficient is equal to zero were as follows.

Table 22: Estimated Regression Coefficients of the categories of SQ15-Recoded and test of hypothesis that coefficient is equal to zero

SQ15-Recoded Category	Estimated Coefficient	t-Value	P-Value
1	2.83	4.131	0.000
2	2.48	3.488	0.001
3	1.43	2.056	0.042

It can be observed from Table 22 that the performance improves gradually from the category SQ15-Recoded = 4 (Never or Do not know) to the categories SQ15-Recoded = 3 (Usually Not), SQ15-Recoded = 2 (Mostly) and SQ15-Recoded = 1 (Yes, Always). Thus, explaining the answers to the questions in detail has a positive impact on the student achievement.

SQ38-Recoded

The SQ38 question was: Being taught by your teacher in small groups. The category SQ38-Recoded = 3 (Once or twice each term or each year) was the reference category. The coefficient for the category SQ38-Recoded = 1 (During most lessons) was not significant, but the coefficient for the category SQ38-Recoded = 2 (Most weeks) was positive with value = 2.06, and the coefficient was highly significant (p-value = 0.00). Thus, “being taught in small groups” is beneficial only if it is done weekly but there is no benefit if it is practiced too often or very rarely.

11. Analyses with IRT Scores

The 2 parameter logistic (2PL) models for each of the 8 tests was estimated using Fractions Book-A (FA) as the reference test. Individual student scores from each test were estimated. The scores for the test FA were scaled from 0 to 1,000 with average of 500 and standard deviation of 100. The test FA was used as reference, and other test scores were estimated with a standard deviation of 100 using the test FA as reference. The average IRT scores for each of the 4 areas of assessment and their aggregate, and the corresponding CVs for each district and the province were estimated. The results are given in Annex 23, Table A-10.0. It can be observed from Table A-10.0 that the district with the lowest achievement was **Umerkot** followed by **Kambar**, and the district with the highest achievement was **Kashmore** followed by **Ghotki**. This result is the same as from the analysis of the classical scores (average percent scores). It should be noted that the CVs of the estimates are somewhat underestimated because the IRT scores are themselves estimates and the uncertainty in the estimated scores has not been accounted for.

The data item SQ38: “being taught by the teacher in small groups” on the student background questionnaire provided interesting result when used as an explanatory variable in the linear regression. Further analyses of data item SQ38 from the student background questionnaire was

undertaken. The item was re-coded by collapsing the original categories 3 and 4. The recoded 3 categories are: 1 = Most lessons; 2 = Most Weeks; and 3 = Once or Twice each term or each year. The estimates of the average IRT scores for the three categories of the item are given in Table 23.

The t-values for the estimated differences between the categories 2 and 1, and the categories 2 and 3 were also computed. The t-values corresponding to the estimated difference of 12.2 between category 2 (Most weeks) and category 1 (Most lessons) was 3.47 (p-value=0.001), and the t-value corresponding to the estimated difference of 28.5 between category 2 (Most weeks) and category 3 (One or twice each term or each year) was 5.54 (p-value=0.000). Thus, these differences were highly significant, and the achievement is the highest for category 2, which is being taught in small groups during most weeks.

Table 23: Average IRT Scores for the categories of SQ38-Recoded

SQ38-Recoded Category	Estimated Average IRT Score	Percent CV
1. Most Lessons	502.3	0.46
2. Most Weeks	514.5	0.50
3. Once or twice each term	486.0	0.93
All Students (Sindh)	505.6	0.32

Item Characteristics Curves

The Item Characteristic Curves (ICCs) were also estimated for each of the items in the 8 tests except for few items for which the bi-serial correlation was less than -0.15. These can be used to revise or replace poor test items. Examples are provided in Annex 24 of some Item Characteristic Curves that have very good discrimination.

12. Constraints, Lessons Learnt and Recommendations

The following constraints, lessons learnt and recommended changes have been identified.

12.1 Item Writing

12.1.1 One of the main challenges for PEACE has been to develop technically sound and appropriately targeted assessment instruments. While PEACE has conducted one round of assessment it is acknowledged that it has far to go to be able to claim that these assessment instruments are “technically sound” and cover the whole of the curriculum identified for testing. Further training workshops are required and PEACE, PITE and BoC staff must practise the skills learnt on a continuous basis to develop mastery

12.1.2 Another challenge has been the requirement to produce sufficient items to be able to conduct the assessment. In some cases it has been impossible during the timeframe to develop items which cover the curriculum content and the skills identified in the assessment frameworks. There is a need for PEACE to plan activities (both institutional and individual) within a realistic time period, with realistic target setting as well as appropriate employment of Technical Assistance (if required).

12.1.3 Delays in the appointment of PEACE additional staff resulted in pressure being placed on a small number of specialists. Also when new PEACE staff were appointed it was difficult for them to be knowledgeable about item development and testing without having hands on experience. This further delayed test development. To assist new staff an induction package should be developed which should contain information such as:

- The names and designations of all PEACE staff;
- The purpose of PEACE;
- The activities conducted by PEACE;
- PEACE achievements;
- Information regarding their roles and responsibilities;
- Technical information regarding their roles and responsibilities.

Time should also be given to one-to-one training of new staff by the appropriate PEACE person.

12.1.4 Another challenge was to ensure the selection of competent subject specialists in the item writing and reviewing workshops. There needs to be further training of these subject specialists to enable them to prepare items for the subject areas at the different cognitive levels identified in the Assessment Framework.

12.1.5 Lack of good English/Urdu/Sindhi language translators. While translations were done there was a lack of rigour in most of the translations resulting in the meaning of some of the items and background questionnaire questions being ambiguous.

12.2 Testing

12.2.1 The pilot and large scale testing may not have been conducted in a standardised manner. This means that the test results might not be based on accurate data. There is evidence from the monitoring of the test administration that the tests might have been conducted with a lack of rigour in some instances for example, not all test administrators used the examples and practice questions found in the students' booklets. For this to be improved requires the test administrators to be selected and trained appropriately and for the monitoring team to be carefully trained and monitoring feedback as well as test administration forms to be utilised to ensure improvement.

12.3 The Sample

The sample was selected using EXCEL. This did not automatically identify discrepancies such as, duplications of SEMIS Codes, replication of schools etc. During the identification of the sample no difficulties were identified. Few checks were conducted by the PEACE staff due to the lack of time available as the testing programme was required to be conducted within strict times with on-the-job training. Also PEACE did not have sufficient capacity or the available budget to enable rigorous monitoring and quality control.

Problems with the sample were identified through information provided by the Test Administrators and through using SAS for analysis.

This lack of implementation of quality procedures resulted in various types of errors being introduced:

- Some sample schools which showed 0 enrollment were found to have sufficient number of students for testing
- Some sample schools which showed high enrollment were found to have 0 enrollment
- Some schools which were required to be split because of large class enrolment, were not identified
- Approximately three schools which had been identified as requiring to be collapsed were found to be too large after collapsing and were then split (this should not have been done)
- Some school not in the identified sample were tested and the data received by PEACE
- Some schools which were in the sample did not send their data to PEACE
- Some schools which had been identified in 2007/08 Census having 0 enrollment and also schools which were identified as closed, were not part of the school sample despite having an appropriate student population

All of these errors resulted in the database being reduced.

The following are the recommendations for improvement:

- SAS software should be used for selecting the sample of schools as SAS code can be easily developed to identify and flag any discrepancies in the data.

- The technique developed during the sampling workshop to identify enrollment discrepancies, where the ratio of the student enrolment at the time of survey and the MOS is lower than 0.5 or higher than 1.5 should be used for further follow-up must be implemented. Where the ratio is found to be <0.5 or > 1.5 additional field checks will need to be made to ensure the reliability of the enrolment of the sampled schools.
- If the measure of size, which is the basis of the sample design, was larger than some threshold value, say 20, and it has changed greatly at the time of test administration it should be investigated. For example, if the MOS was 20 and it changes to 50, or if the MOS was 100 and changes to 20, these are serious discrepancies and should be investigated.

12.3 The Assessment Instruments

There was no full-time In-Page composer available in PEACE for producing the tests. This resulted in delays in the composing and the setting of the tests, guides and background questionnaires. Due to this and the time constraint in which all the activities had to be completed there was a delay in the printing of the materials;

The Mathematics tests had some misprinting and due to the time constraints and little or no monitoring of the instruments no corrections were made. This had implications for the statistical analysis of the data.

No check list was provided to the markers and coders to help them to self monitor their own data entry on to the marking and coding sheets.

It is essential that a full time In-Page composer is available for test development in PEACE. Improved quality checks need to be made at all stages of test development.

12.4 Test Administration

Some of the **difficulties** identified in the test administration were as follows:

- Some of the district focal persons and test administrators did not always appreciate the need for the assessment to be conducted in a rigorous manner;
- Test administrators did not successfully demonstrate the example questions and practice test questions in the test booklets to familiarize the students with the test methodology;
- Test administrators did not always follow the procedures given in the guidance booklet.
 - ✓ Test administrators found the use of the random number table (used to identify 10 students in a class of more than 10 students) as well as the skip interval difficult to understand and practice
 - ✓ There was also a lack of understanding of the methodology for entering the correct information for “split” schools

- The school enrollment was not recorded during the test administration for some of the sampled schools

There is a need for improvements in the training of the focal persons and test administrators. There is also a need to develop improvements in the monitoring of test administration.

12.5 Data Entry and Analysis

No appropriate Guidelines or training was provided to data entry personnel and this resulted in mistakes in the entry of the SEMIS code, split schools, gender, location information.

There was a delay in the statistical analysis due to the delay in entering the assessment data.

12.6 Logistics

- Time constraint. It was difficult to ensure that all the tasks were completed on time in a rigorous manner;
- The transportation of the assessment materials to arrive in the districts to enable training and testing to take place was difficult. The provinces and areas also experienced difficulty in ensuring that the assessment instruments reached all the sample schools on time for testing. This was especially true in the remote areas of Sindh;
- The collection of the assessment instruments after testing had taken place was also difficult especially in remote areas;
- The test administration training in many provinces was hampered due to the delay in the arrival of the training materials
- There was a lack of appreciation for the need for the assessment materials to be kept in a secure location and for all the assessment instruments to be returned to the focal persons and then PEACE;
- The approved flat and uniform rates of transportation allowance did not take into account the availability or otherwise of transport facilities. Fares vary according to the area, being higher in difficult areas such as mountains and desert areas;
- Communication between the test administrators, PEACEs, and focal persons was often difficult; focal persons/PEACE were often unable to assist in the solving of day to day problems on their own.

12.7 Management

- Some of the staff who had been trained were not available to conduct the assessment as they had been transferred or posted to another post;
- The monitoring tool was too simple for effective monitoring to take place– it required only a yes or no answer;
- The coordinators responsible for making payments to the field staff did not pay the staff promptly;
- No honoraria were provided to Head Teachers, sometimes resulting in a lack of cooperation.

12.8 Staffing

- There was not sufficient staff in the NEAS or the PEACEs or AEACs to ensure that the assessment was conducted in an efficient and timely manner. Many of the staff were assigned multiple tasks due to this constraint;
- Some of the Executive District Officers (schools) demonstrated little interest in the ongoing assessment process: they did not respond to correspondence in connection with the nomination of test administrators and the test administrators were not informed in sufficient time to attend the training.

12.9 Equipment

- There was insufficient equipment (computers, photocopiers) available at the training centres for the duplication of materials for training purposes;
- In marking and coding the scripts there was not sufficient space available for storing the assessment materials.

12.10 Recommendations

- There is a need for the employment of a full time In-Page composer to enable the production and printing of the 2010 assessment instruments to be conducted in a timely fashion;
- An independent editor is required to review the tests before they are printed and also after the first sample printing has been completed to ensure that there is no misprinting;
- The training and assessment instruments should be distributed to the districts in sufficient time for training and the administration of the assessment instruments to take place efficiently;
- There is a need to emphasize that the assessment instruments should be kept securely and that all the assessment instruments, whether used or unused, should be returned to the district focal persons and then on to PEACE;
- From the 2009 experience monitors should be trained to address day to day problems;
- Quality Assurance Guidelines have been developed for Monitoring the Marking and Coding;
- It is hoped that through dissemination all the stakeholders will understand the importance of conducting provincial assessments and assist in the activity;
- Marking and coding of the assessment instruments should be centralized at one place such as Hyderabad and monitored by persons who have attended the analysis workshops to enable scoring errors to be reduced.