



**Education & Literacy Department
Government of Sindh**

**Sindh Students Assessment
Results Report
2011
Language Grade 4**

**Provincial Education Assessment Centre (PEACE)
Bureau of Curriculum and Extension Wing Sindh, Jamshoro**

Coordinated by: Reform Support Unit Sindh, Karachi

May 2013

Acknowledgement

The Education and Literacy Department acknowledges the work undertaken by:

PEACE BC&EW Sindh Jamshoro in developing, designing, administering, analyzing and report writing and dissemination for Sindh Students Grade 4 language Assessment 2010.

Focal persons in the districts who manage the training for test administrators, distribution and collection of the assessment instruments

Test administrators arduous work in administering the test.

The students who took the tests

The head teachers, teachers and students who took the background questionnaires

The RSU who provide policy guidelines and supported the whole activity

EU SER-TA who provided the technical support

BC&EW Sindh Jamshoro for the dissemination of results Sindh Students Grade 4 language Assessment 2010.

DRAFT

Table of Contents		Page
	Acknowledgement	i
	Table of Content	ii
	List of Tables	iii
	List of Figures	iii
	Acronyms	iv
	Executive Summary	v
1.	Introduction	1
2.	Sindh Province Testing Model	2
3.	Language Assessment 2010	3
4.	Assessment Instruments	4
5.	Assessment Sample	5
6.	Test Administration, Monitoring and Coding, Data Entry and Cleaning	6
	6.1 Test Administration	6
	6.2 Test Monitoring	7
	6.3 Coding and Data Entry	8
7.	Item and Background Data Analyses	9
8.	The Results	10
	8.1 Understanding the Assessment Results	10
	8.2 Findings	11
	8.3 Background Questionnaires	23
9.	Constraints	26
10.	Lessons Learnt	28
11.	Implications	28
12.	Recommendations	29
	Annexes	30
	1. List of Provincial Working Group members	31
	2. Language Test Specifications	32
	3. Test Instrument Development	33
	4. Developing the Provincial Test Sample	36
	5. The Grade 4 Language Analyzed Provincial Sample	42
	6. Coding, Data Entry and Cleaning	44
	7. Examples of Test Items	47
	8. Glossary of Terminology	51

List of Tables

1.	Number of Schools and Test Administrators	6
2.	Sindh Province Mean Language Score	11
3.	Student Language Achievement by Location	12
4.	Student Language Achievement by Gender	13
5.	Urdu/ Sindhi Language Mean Score by Districts	14
6.	Districts Reading and Writing % Mean Score	15
7.	Reading Mean Score According to Medium of Instruction	16
8.	Reading Mean Score According to Reading Contexts	17
9.	Reading Mean Score According to Aspects Tested	18
10.	Student Achievement According to Writing Aspect Tested	19
11.	Student Achievement According to Writing Criteria	20
12.	Student Writing Results by Gender and Medium of Instruction	21
13.	Student Writing Mean Score by Location and Medium of Instruction	21

List of Figures

1.	District Wise Coverage of Grade 4 Schools Sample	5
2.	Sindh Province Mean Language Sindhi/ Urdu Score	11
3.	Student Language Achievement by Location	13
4.	Student Language Achievement by Gender	13
5.	Urdu/ Sindhi Language Mean Score by District	15
6.	Districts Reading and Writing Results	16
7.	Mean Score According to Medium of Instruction	17
8.	Reading Mean Score According to Reading Contexts	18
9.	Reading Mean Score According to Aspects Tested	19
10.	Student Achievement According to Writing Aspects Tested	20
11.	Student Achievement According to Writing Criteria	21
12.	Student Writing Mean Score by Gender and Medium of Instruction	21
13.	Student Writing Mean Score by Location and Medium of Instruction	22

Acronyms

BC&EW	Bureau of Curriculum and Extension Wing Sindh
BQ	Background Questionnaire
DFP	District Focal Person
EDOE	Executive District Officer Education
EU	European Union
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
MCQ	Multiple Choice Question
MOS	Measure of Size
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
RA	Reading Book A
RB	Reading Book B
RSU	Reform Support Unit
SAS	Statistical Analysis System
SEMIS	Sindh Education Management Information System
SERP	Sindh Education Reform Programme
SER-TA	Sindh Education Reform Technical Assistance
STEDA	Sindh Teacher Education Development Authority
TA	Test Administrator
WA	Writing Book A
WB	Writing Book B

Executive Summary

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

To support improvements in learning quality a Language 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 Language 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 Sindhi and Urdu Languages
- To use information regarding students' attitude to Sindhi and Urdu language and language teaching to improve the quality of education and students' learning.
- To use information regarding students' attitudes to Language and Language teaching to improve the quality of teacher training and teachers' classroom practice.

Grade 4 Language reading and writing were assessed. The survey was planned with 37,770 students in all districts of the province. Background questionnaires were given to Head Teachers, teachers and students in the sampled schools to identify the relationship between, student achievement and the factors influencing their achievement.

The main results of the survey are as follows:

- The overall **language** mean score of students was 44.0%;
- Eleven out of the 23 districts performed above the mean performance of the rest of the districts in the province;
- Students achieved an overall **reading** score of 46.2%
- Students in rural areas (46.8%) achieved a higher reading comprehension score than students in urban areas (44.0%);
- Both girls and boys achieved a similar overall result in reading (46.2%)
- Students performed best in answering questions on a story (48.4%) and weakest on answering questions on poetry (44.1%)
- Students demonstrated their best achievement in answering questions which required them to demonstrate a general understanding of what they had read (44.3%) while they were weakest in answering questions on reader-text connections (35.2%)
- Students achieved an overall **writing** score of 41.71%
- Students performed best in narrative writing tasks (47.7%)
- Students showed weak achievement in the use of punctuation (15.86%)
- Students performed best at identifying the meanings of words (56.6%) and were weakest in their responses to questions about irony and words having opposite meanings (16.3%)
- Girls (43%) achieved a higher mean score than boys (40.8%) in the writing tasks;

- Students in urban areas (42.8%) performed better in writing tasks than those in rural areas (41.3%)

From this study it can be seen that much work is required to improve the teaching and learning of reading and writing skills. The results of these tests are found in detail in Section 11.

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; some strategies for the improvement of reading and writing are found in the Feedback Report;
- Interpret the Language 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 Language 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.

A complete Technical Report is available from the RSU and PEACE.

1. Introduction

The improvement of the quality of education requires a multi directional approach in order to improve the effectiveness of the learning process by improving teaching techniques, supplying better learning aids, motivating students to attend school etc. Student assessments which are credible and objective play a pivotal role in this process by providing critical feedback on what and how well students are learning¹.

PEACE, Sindh was initially developed as a unit to support the development of national assessments conducted by National Education Assessment System (NEAS). Between 2004 to 2008, four national 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 statistical analysis, report writing and the dissemination of results through this development.

Further training and support has been provided to PEACE under the SERP by EU technical assistance. 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.

On the basis of the National Assessment findings, Sindh PEACE has developed a system of diagnostic assessment to find out and diagnose where students have difficulties and where support is required to improve their achievements. This has involved the development, conduct and the statistical analysis of a large sample of student data.

Mathematics, languages, science and social studies have been identified as subjects to be tested by PEACE. Mathematics was chosen to be tested first. PEACE has already reported on the Mathematics achievement of the students in 2010. Reports regarding the mathematics achievements can be obtained from the Reform Support Unit and PEACE. Through these assessments of Mathematics and Language learning (using assessment to make sound judgments about learning and school effectiveness) sound judgments can be made about learners' achievements. This assessment also supports assessment for learning (using assessment to support classroom teaching) and assessment as learning (using assessment to promote autonomy in learning) by ensuring that standards are understood and shared by teachers, the learners themselves, their parents and other adults who are either directly engaged in or are supporting the learning process.

The process of the preparation of the report of the 2010 Grade 4 Language Provincial Assessments has involved all key stakeholders. An Assessment Working Group was constituted and assigned responsibility for reporting in detail on one or more sections which would ultimately form the report. These contributions were subsequently compiled into a first draft at the RSU. A list of the members of this group is found in Annex 1. The draft was further refined and additional material provided by PEACE has been incorporated into the 2010 Grade 4 Language Assessment Report.

Further information regarding any aspect of the provincial tests of Language can be obtained from PEACE, BC&EW.

¹ The World Bank Report, December 2008

2. 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 and cannot do is stated. This requires descriptions of performance as found in the Language 2006 National Curriculum for Pakistan. 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 the Grade 4 Language, 2006 National Curriculum and identifying the competencies to be tested. From this two domains in reading, were identified for testing, namely, context and **cognitive domains** and three types of writing were identified for testing (narrative, informative and persuasive writing). Test specifications were written to develop test items and writing prompts to assess these specific reading domains and writing skills, according to the weightage indicated in the National Curriculum.

Test items and writing prompts 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 selected 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.

3. Language Assessment 2010

The Provincial Education Assessment Centre (PEACE) was established in 2002 to provide objective information on student performance to policy makers, the Bureau of Curriculum, teacher training institutions and officers at provincial, district, taluka, Union Council and school levels and to provide stakeholders with an evaluation of the condition and progress of education. 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.

Sindh Province Grade 4 Language Assessment, 2010 is a provincially representative and continuing assessment of what Sindh Province students know and can do in various aspects of Language. The first Language Provincial Assessment was conducted in all the districts of Sindh in 3777 schools (Primary Sample Units) and with 28,672 students who answered 61,554 tests.

The Language Assessment Framework is based on a consensus regarding desirable elements of language education against which student attainment is to be measured. As mentioned previously, the purposes of language identified in the Language National Curriculum are all to be regarded as important. The language assessment should enable students to demonstrate achievement through the use of paper and pencil tests. For this reason listening and speaking are not part of the Language framework. These two skills however are very important and students should be provided with classroom experiences to enable them to listen and communicate through language

PEACE along with working school teachers, BC&EW, PITE and GECE staff developed the Language Assessment Framework in 2008. This framework identified the two skills of reading and writing in the National Curriculum to be tested. 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 reading and writing skills and aspects of reading and writing. Each content domain has several topic areas.

The test specifications on which the Grade 4 Language assessments were based are found in Annex 2.

4. Assessment Instruments

To measure the students' learning achievements as well as getting factors which affect the quality of education in Sindh Province two different assessment instruments were developed to test Grade 4 students in 2010. Two types of instruments developed were:

- Language Diagnostic tests in Reading and Writing;
- Background Questionnaires for head teachers, teachers and students.

Language Diagnostic tests were developed on the basis of the National Curriculum, 2006. The reading and writing tests specification required:

- Using the Languages (Sindhi and Urdu) 2006 National Curriculum to identify the content areas to be tested;
- Identifying from the National Curriculum the weightage to be given to each area to be tested;
- Identifying the aspects of reading (forming a general understanding, developing interpretation, making reader-text connections and examining the content and structure of a text) as well as the contexts for reading (reading for literary experience and reading for information) to be tested.
- Identifying the aspects of writing as narrative writing – telling a story, informative writing (informing the reader); persuasive writing (persuading the reader). Knowledge about language (Grammar and punctuation) was also to be tested. This would include the aspects of vocabulary, grammar and sentence structure.

Some examples of the Language diagnostic test items are found in Annex 7.

Background Questionnaires were developed for Head Teachers, Teachers and Students. These questionnaires contained questions to identify the effects of, for example, teacher qualifications on student results; the effect of giving/not giving homework on student achievement; the effect of living in a rural area as compared with an urban area, and, student attitudes regarding Language and Language teaching.

The Language diagnostic tests were piloted in 2009 on a sample of 1150 students. These tests were then marked, coded and entered into a data base. From the results of the statistical analysis of the pilot items (using ITEMAN software) items were selected and additional items developed where required. Formats for large scale testing were developed and the tests were administered in April 2009 to a provincial sample of 28,866 Grade 4 students to establish a baseline of achievement and to identify specific problem areas in learning for Grade 4 students in Language.

Additional information on the development of the assessment instruments is found in Annex 3.

5. Assessment Sample

It was proposed that in 2008 a representative sample (15% of the Grade 4 population) of approximately 40,000 Grade four students (3777 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). The coverage of the provincial sample for Grade 4 Language is found to be good. The actual sample used was 28,684 students and 3777 PSUs. This was as a result of verification of the status of the schools by District Officials. The final sample used for analysis was 3,746 schools. This was a result of the data cleaning exercise conducted by PEACE.

The results obtained from this assessment provide Sindh Province with a benchmark for all future results to be judged. The sample was selected according to district; rural/urban; male/female categories covering the whole province. More detail regarding the development of the sample is found in Annex 4.

A map showing the district wise coverage of the Grade 4 school sample is found below.

Fig 1: District wise coverage of the Grade 4 School Sample



More detailed information is found in Annex 5.

6. Test Administration, Monitoring and Coding, Data Entry and Cleaning.

6.1 Test Administration

The tests were administered in April, 2010 by Test Administrators in the districts according to instructions in the Test Administration Guideline Booklet. The teachers were trained by Master Trainers who had previously received training by PEACE. The PEACE staff monitored the test administration. One hundred master trainers from all districts were trained by PEACE, at four centres, Hyderabad, Mirpurkhas, Larkana and Sukkur. These master trainers trained test administrators throughout the Province. The number of test administrators trained was as follows:

Table 1: Number of Planed Schools and Test Administrators

S.No	District	No. of Schools	No. of Test Administrators
1	Badin	172	172
2	Dadu	201	201
3	Ghotki	188	188
4	Hyderabad	153	153
5	Jacobabad	142	142
6	Jamshoro	111	111
7	Kambar-Shahdadkot	171	171
8	Karachi	278	278
9	Kashmor@ Kandhkot	138	138
10	Khairpur	210	210
11	Larkana	188	188
12	Matari	120	120
13	Mirpurkhas	158	158
14	Nausheroferoz	183	183
15	Sanghar	201	201
16	Benazirabad	166	166
17	Sukkur	153	153
18	Shikarpur	147	147
19	Tando Allah Yar	107	107
20	Tando Muhammad Khan	96	96
21	Tharparkar	167	167
22	Thatta	174	174
23	Umarkot	142	142
	TOTAL	3766	3766

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

- Some of the 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 always use the examples in the test booklets to familiarize the students with the test methodology;
- Test administrators did not always follow the guidance given in the guidance booklet.
 - ✓ Test administrators found the use of the random number table (used to identify 10 students in a class which had more than 10 students) as well as the skip interval difficult to understand and practices
 - ✓ 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

6.2 Test Monitoring

Monitoring was undertaken in two areas:

- Monitoring the Test and BQ Administration, and,
- Monitoring the marking and coding of the assessment instruments

6.2.1 Monitoring the Provincial Assessment Administration

The main objective of monitoring the provincial assessment of the test administration was to ensure the validity of the Assessment data. It is important that all aspects are standardized, including the administration of the assessment instruments.

The monitors consisting of 23 district focal persons PEACE, BC&EW, PITE and EDOs were involved in the overall monitoring. The monitors monitored the assessment activity and reported back to PEACE on how well the test administrators followed the guidelines given during the test administration training.

6.2.2 Monitoring the Marking and Coding of the Assessment Instruments

On the recommendation of the previous mathematics survey to ensure quality assurance, centralized coding and marking of all the scripts was undertaken in Hyderabad. Monitoring of this activity was undertaken by persons involved in the analysis workshop. Instructions regarding the coding and marking were given to the personnel involved.

The responsibility of the monitors was to:

- Ensure that the coding and marking was conducted in an efficient, fair and transparent manner;
- Provide support to those coders and marking personnel who were having some difficulties
- Identify coding and marking discrepancies and correct them where possible;
- Take a 10% sample of the coding sheets to identify the reliability of the coding and marking;
- Ensure that the correct code was being used on the scoring sheets;
- Ensure that the assessment instruments were complete and returned to PEACE when the coding and marking was completed.

6.3 Coding and Data Entry

For the 2010 Language 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 and for writing, rubrics were developed.

Checking the data was an onerous a task so it was not possible to check every single sheet. The data of two students 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 one out of 10 students on each scoring sheet in Hyderabad. Where discrepancies were identified these were rectified where possible or the data was discarded due to its lack of reliability.

The cleaned data was then entered into the SAS program ready for data cleaning. More detail on the data entry and cleaning can be found in Annex 6.

DRAFT

7. Item and Background Data Analyses

The analysis of test results to provide valid and reliable information is a complex task. Data analyses involved following phases:

- Phase I included Classical item and reliability analysis of the pilot test items for the selection of test items for the large-scale testing;
- Phase II consisted of the cleaning of the large-scale testing scores and background data information using SAS
- Phase III consisted of IRT analysis using Bilog-MG.
- Phase IV consisted of the use the WesVar program to analyze the data.

The following is a description of software used for each phase of analysis.

Phase I: SAS PROGRAM

This was used at the stage of sample selection and is helpful in basic arrangements of sampling frames, calculations of sampling intervals, generating random number seeds and sorted lists of sampled schools.

Phase II: ITEMAN

Item analysis of the pilot test items was carried out using **ITEMAN**. This is “Classical item analysis” for the purpose of item selection. This program provided information on two parameters; item difficulty and item discrimination and was used in item selection along with information on the performance of distractors for each item.

Phase III: SAS Program

SAS program was used to clean all the data from the large-scale testing and Background Questionnaires. The SAS program provides a complete, comprehensive set of tools that meets the data analysis needs required for provincial assessment.

Phase IV: Bilog-MG3 Program

Bilog-MG3 was used to provide Item Response Theory (IRT) information. It models the response of each student of a given ability to each item in the test. Bilog-MG3 converts students' raw scores on a test or versions of a test to a common scale that allows for a numerical comparison between students. The PEACE testing programs used multiple versions of a test, and the scale is used to control slight variations from one version of a test to the next. Scaled scores are particularly useful for comparing test scores over time, such as measuring semester-to-semester and year-to-year growth of individual students or groups of students in a content area.

Phase V: WesVar Data Analysis

WesVar was used for the analysis of the survey data and for linear and Logistic regression analysis.

8. The Results

8.1 Understanding the Assessment Results

PEACE uses widely accepted statistical standards in analyzing the data. More detailed statistical information can be found in the comprehensive technical report produced by the Government of Sindh.

Results are presented as mean percentage scores. The achievement of students is discussed in Section 8.2 according to the following topics:

- Overall student achievement
- Student Achievement by Location
- Student Achievement by Gender
- District results
- Differences between student achievement in each district and the rest of the districts
- Student overall achievement according to Reading Contexts
- Student overall achievement according to Reading Aspects
- Student overall achievement according to Writing Aspects
- Student achievement According to Writing Criteria
- Student overall achievement according to Gender and Location
- Student overall achievement according to Medium of Instruction
- Background Questionnaire Findings

To make the scores meaningful and to establish a relationship between student achievement and through the various variables, classical analysis and significance levels are explained below.

8.1.1 Classical Analysis

Classical analysis was used to identify “good” test items from the pilot testing. The items were Classical as “good” if they demonstrated good reliability and validity and if they were able to discriminate between students of different abilities. Classical analysis was also conducted to enable stakeholders to relate more easily to the actual achievement scores of the students as related to the competencies tested. This Classical analysis is presented in percentage form for the four mathematical areas tested.

8.1.2 Significance Levels

To check whether differences in reported scores could have occurred by chance alone, significance tests are reported. A probability where is < 0.05 means that the difference could occur by chance alone in only 5 out of 100 students; where it is < 0.01 the difference could occur by chance alone in only 1 out of 100 students (significant difference) and where it is 0.00 there is a highly significant difference.

8.2 Findings

Two sets of findings were obtained from the assessment namely:

- Assessment Test results
- Background Questionnaire results.

8.2.1 Assessment Test Results

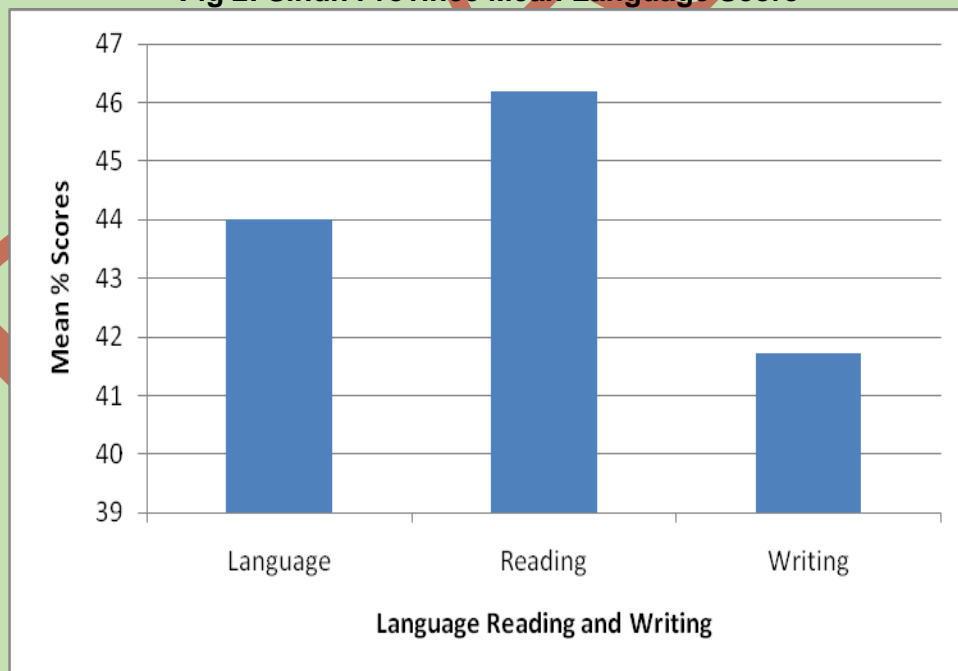
8.2.1.1 Overall Student Language Achievement

This report presents the Sindh Provincial results of the assessment in Language (reading and writing). On the 2010 Grade 4 Language tests students achieved a provincial percentage mean score of 44% in Language, having a mean score of 46.2% in reading and a mean score of 41.71% in writing

Table 2: Sindh Province Mean Language Score

MEAN SCORE		
LANGUAGE	READING	WRITING
44.00%	46.2%	41.71%

Fig 2: Sindh Province Mean Language Score



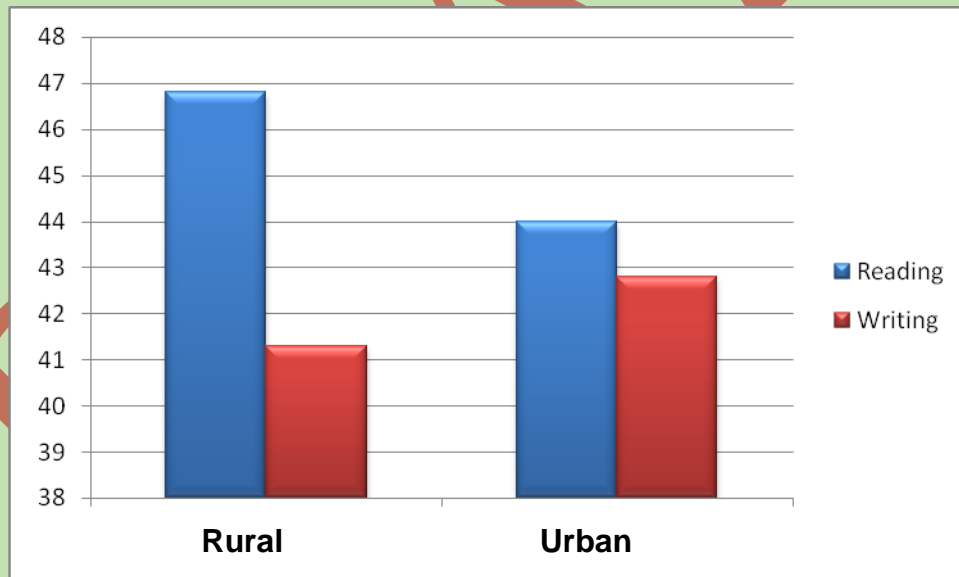
8.2.1.2 Overall Student Language Achievement by Location

It was noted that students in rural areas performed better in reading as compared with students in urban areas. The difference between rural and urban student performance is highly significant in reading and their performance in writing shows a significant difference.

Table 3: Student Language Achievement by Location

ASPECT	LOCATION		PROBABILITY	Significant / Not Significant
	RURAL	URBAN		
Reading	46.8	44.0	0.00	Highly Significant
Writing	41.3	42.8	0.01	Significant

Fig. 3: Student Language Achievement by Location



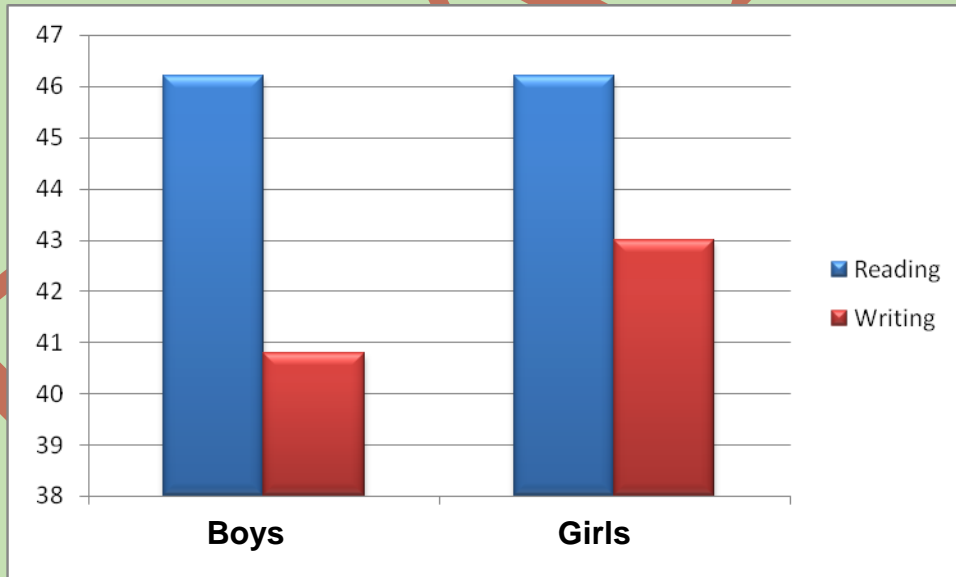
8.2.1.3 Overall Student Language Achievement by Gender

In reading there was no significant difference between the achievement of boys and girls in Sindh Province. In writing the difference between boys and girls' performance was highly significant with girls' achievement of 43.0% being higher than boys (40.8%).

Table 4: Student Language Achievement by Gender

ASPECT	GENDER		PROBABILITY	Significant / Not Significant
	BOYS	GIRLS		
Reading	46.2	46.2	0.98	Not Significant
Writing	40.8	43.0	0.00	Highly Significant

Fig. 4: Student Language Achievement by Gender



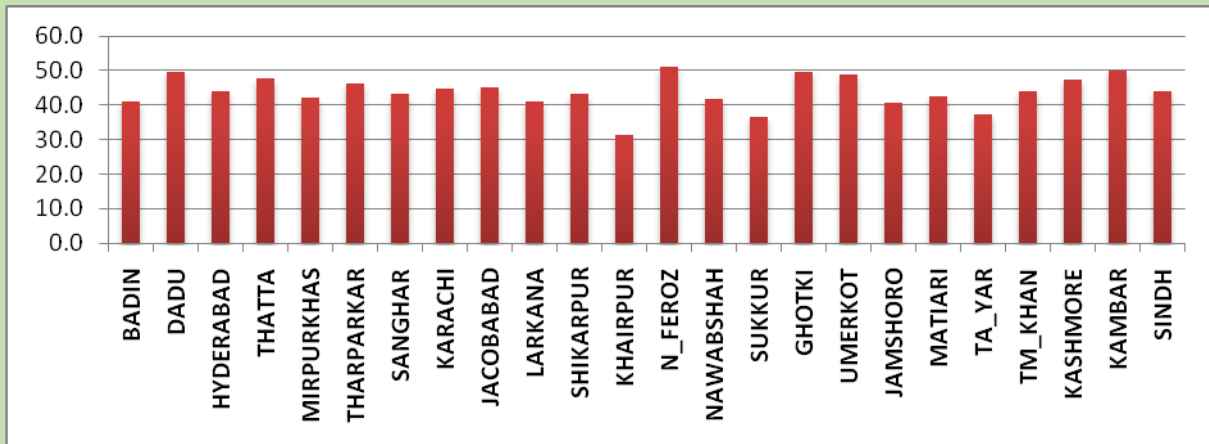
8.2.1.4 Students' Language Mean Score by District

Eleven out of the 23 districts performed above the mean performance of Sindh province as a whole. The district with the highest score in language was Naushero Feroze (51.3%) while the weakest performance was found in Khairpur (31.4%)

Table 5: Language Mean Score by District

Language Mean Score District		
District_ID	District Name	Mean Score
1	BADIN	41.0
2	DADU	49.5
3	HYDERABAD	44.1
4	THATTA	47.8
5	MIRPURKHAS	42.0
6	THARPARKAR	46.4
7	SANGHAR	43.2
8	KARACHI	44.8
12	JACOBABAD	45.2
13	LARKANA	40.9
14	SHIKARPUR	43.1
15	KHAIRPUR	31.4
16	N_FEROZ	51.3
17	NAWABSHAH	41.8
18	SUKKUR	36.5
19	GHOTKI	49.7
20	UMERKOT	49.0
22	JAMSHORO	40.6
23	MATIARI	42.7
24	TA_YAR	37.2
25	TM_KHAN	43.8
26	KASHMORE	47.4
27	KAMBAR	49.8
Language Mean Score		44.0

Fig 5: Language Mean Score by District



8.2.1.5 Student Achievement as Compared with all Other Districts of the Province

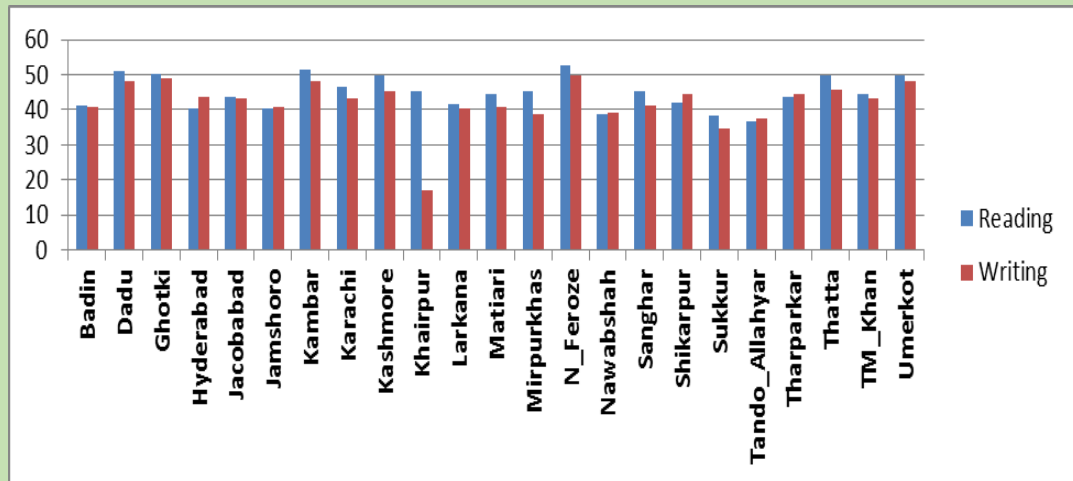
Students from seven districts (**N-Feroze, Kambar, Dadu, Ghotki, Umerkot, Thatta and Kashmore**) achieved a higher percent score in reading as compared with all the other districts of the province; in writing **N-Feroze, Ghotki, Kambar, Umerkot, Thatta, Dadu, Kashmore and Shikarpur** achieved a higher percent mean score as compared with the other district of province

Table 6: Districts' Reading and Writing % Mean Score

District ID	District Name	Reading % Mean Score	Writing % Mean Score
1	Badin	41.2	40.8
2	Dadu	50.9	48.0
19	Ghotki	50.4	48.9
3	Hyderabad	40.4	43.5
12	Jacobabad	43.7	43.1
22	Jamshoro	40.2	40.9
27	Kambar	51.4	48.3
8	Karachi	46.5	43.2
26	Kashmore	49.7	45.2
15	Khairpur	45.5	17.2
13	Larkana	41.5	40.2
23	Matiari	44.6	40.8
5	Mirpurkhas	45.2	38.8
16	N_Feroze	52.6	49.9
17	Shaheed Benazirabad	38.9	39.3
7	Sanghar	45.2	41.2
14	Shikarpur	41.9	44.4
18	Sukkur	38.3	34.6
24	Tando_Allahyar	36.8	37.6
6	Tharparkar	43.7	44.3

4	Thatta	49.7	45.9
25	TM_Khan	44.3	43.3
20	Umerkot	49.8	48.2

Fig 6: Districts' Reading and Writing Results



8.2.2 Students' Reading Achievement

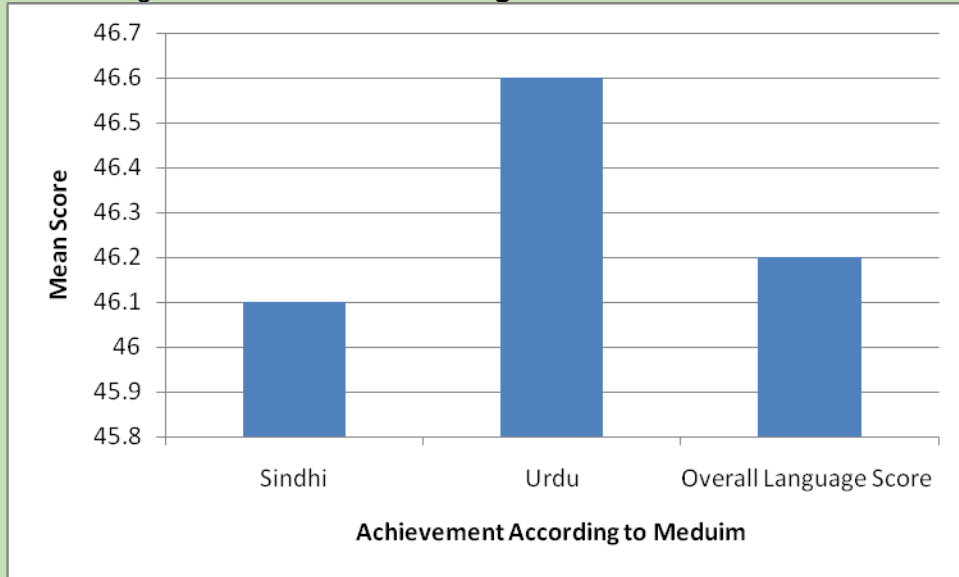
8.2.2.1 Reading Mean Score According to Medium of Instruction

The medium of instruction appeared to have little effect on the scores of the students.

Table 7: Reading Mean Score According to Medium of Instruction

MEDIUM	% MEAN SCORE
Sindhi	46.1
Urdu	46.6
Reading Mean Score for Province	46.2

Fig 7: Mean Score According to Medium of Instruction



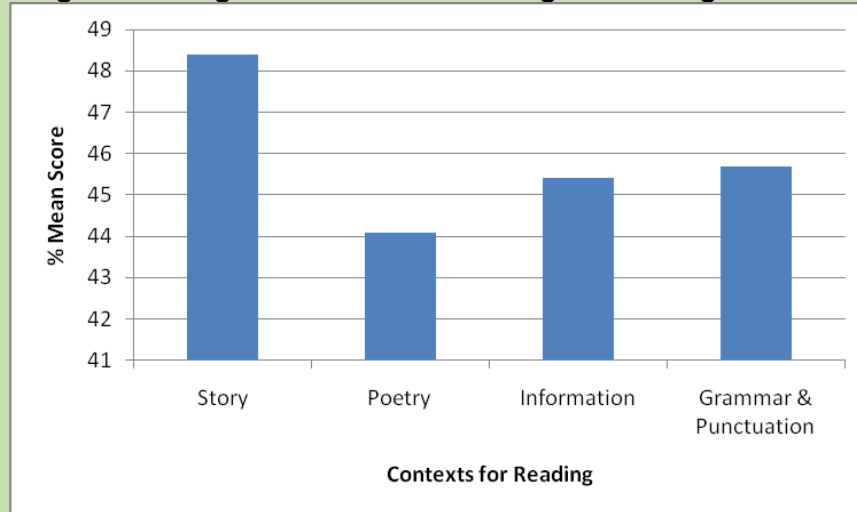
8.2.2.2 Reading Mean Scores According to Reading Contexts

Three forms of reading contexts as well as grammar and punctuation were used to test students' reading comprehension skills. Students found reading comprehension tasks difficult with all scores below 50%. Poetry was found to be the most difficult (44.1%) with students achieving the highest mean score for reading and answering questions on a story (48.4%).

Table 8: Reading Mean Score According to Reading Contexts

READING CONTEXTS	% MEAN SCORE
Reading for Literary Experience (Story)	48.4
Reading for Literary Experience (Poem)	44.1
Reading for Information	45.4
Grammar and Punctuation	45.7

Fig 8: Reading Mean Score According to Reading Contexts



**Table 8: Reading for Literary Experience (Story)
Mean Score by Gender and Medium**

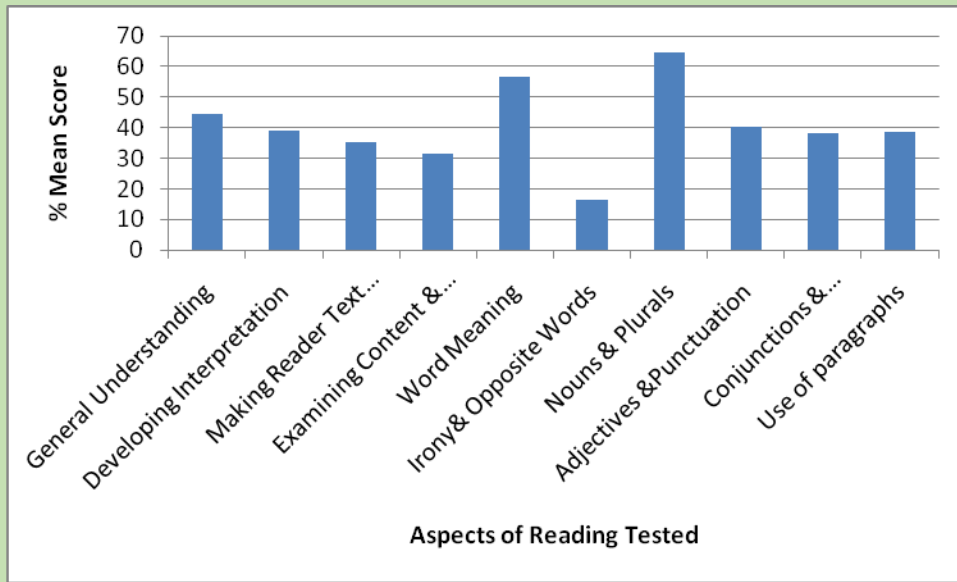
8.2.2.3 Reading Mean Scores According to Reading Aspects Tested

Ten aspects of reading were tested. Students found the questions on irony and opposite words the most difficult (16.3%) followed by questions on examining content and structure (31.5%). Students achieved 64.4% on questions about nouns and the plural of words and 56.6% on questions about word meanings. The aspects which students found most difficult were the aspects which required the students to think and respond to the questions asked – general understanding, interpretation, making reader-text connections and the content and structure of a text.

Table 9: Reading Mean Score According to Reading Aspects Tested

READING ASPECTS TESTED	% MEAN SCORE
General Understanding	44.3
Developing Interpretation	38.9
Making Reader Text Connection	35.2
Examining Content and Structure	31.5
Word Meaning	56.6
Irony and Opposite Words	16.3
Nouns and Plurals	64.4
Adjectives and Punctuation	40.3
Conjunction and Alphabetical Order	38.1
Paragraphs	38.5

Fig 9: Reading Mean Score According to Reading Aspects Tested



8.2.3 Students' Writing Achievement

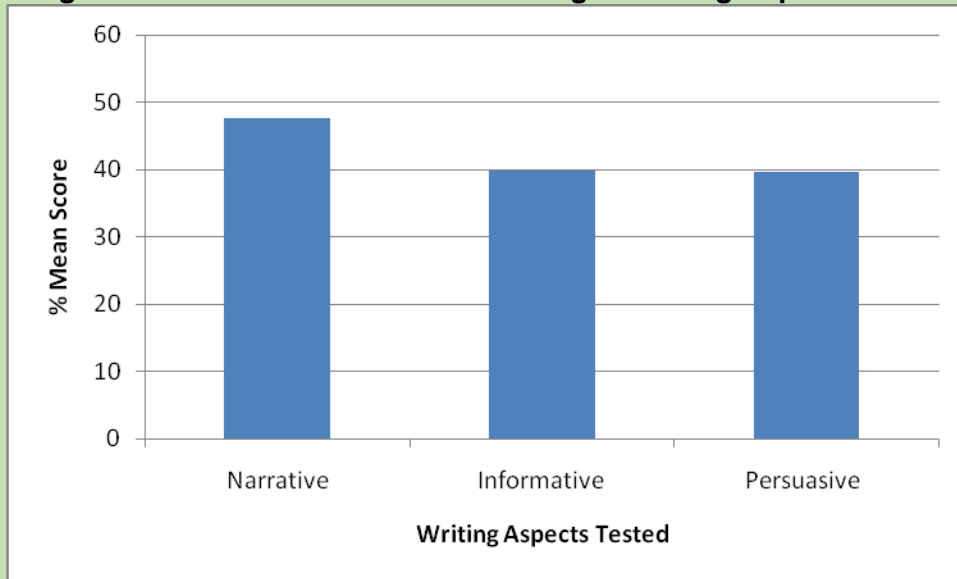
8.2.3.1 Student Achievement According to Writing Aspects Tested

Students found the writing tasks difficult, especially the informative and persuasive writing tasks.

Table 10 Student Achievement According to Writing Aspect Tested

WRITING ASPECT TESTED	% MEAN SCORE
Narrative	47.7
Informative	39.9
Persuasive	39.6

Fig 10: Student Achievement According to Writing Aspect Tested



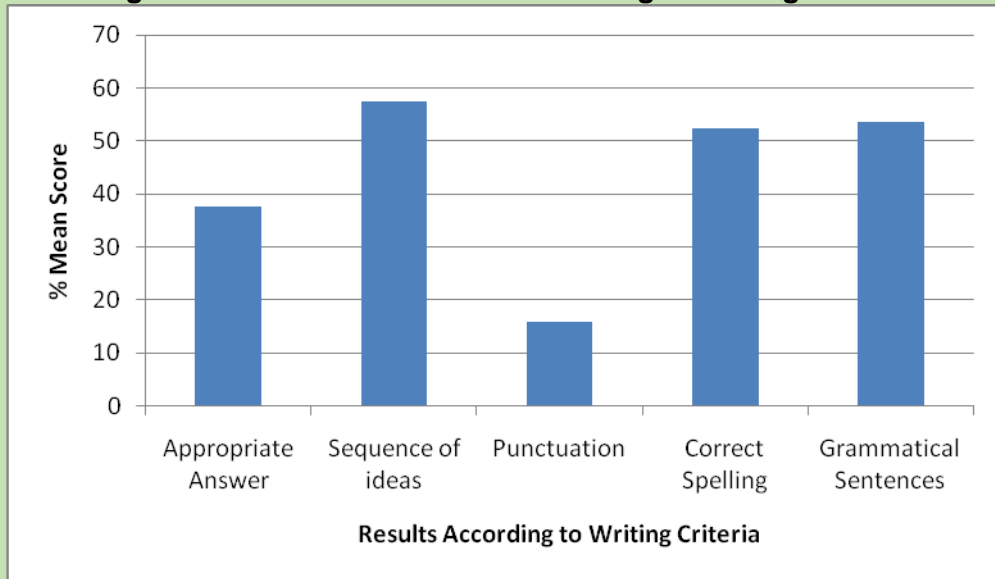
8.2.3.2 Student Achievement According to Writing Criteria

The results in student achievement of all areas of writing criteria were weak. From an analysis of the achievement of students according to their achievement of writing criteria it can be seen that students achieved the highest scores in the sequencing of ideas, writing correct grammatical sentences and using correct spelling.

Table11 Student Achievement According to Writing Criteria

WRITING CRITERIA	% MEAN SCORE
Appropriate Answer	37.53
Sequence of ideas	57.44
Punctuation	15.86
Correct Spelling	52.44
Grammatical Sentences	53.63

Fig 11: Student Achievement According to Writing Criteria



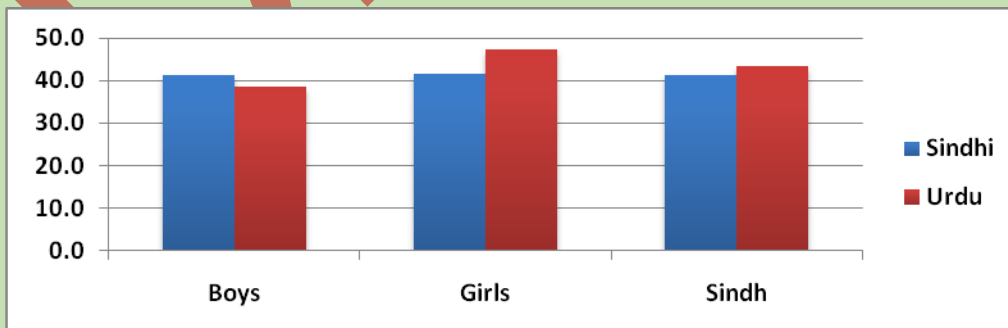
8.2.3.3 Students' Writing Results by Gender and Medium of Instruction

For boys there was a highly significant difference in overall writing achievement in the Sindhi and Urdu medium tests; for Girls there was a highly significant difference in achievement and overall the difference between Sindhi and Urdu performance is highly significant.

Table 12: Students' Writing Mean Score by Gender and Medium of Instruction

Gender	Sindhi	Urdu	Prob:	Significant/Not Significant
Boys	41.2	38.6	0.01	Highly Significant
Girls	41.6	47.4	0.00	Highly Significant
Sindh	41.3	43.5	0.01	Highly Significant

Fig 12: Students' Writing Mean Score by Gender and Medium of Instruction



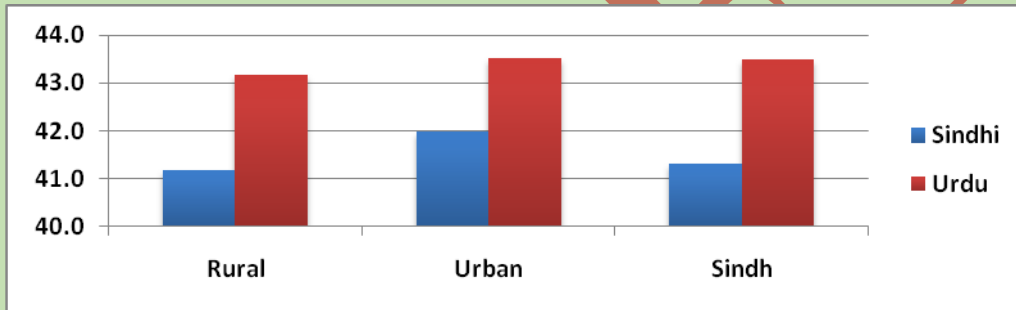
8.2.3.4 Language Results for Writing by Location and Medium of Instruction

In Rural areas there was no significant difference in the scores in the Writing Sindhi/Urdu medium tests and also in Urban areas there was no significant difference. Overall in Sindh the difference between Sindhi/ Urdu performance is highly significant.

Table 13: Writing Mean Score by Location and Medium of Instruction

Location	Sindhi	Urdu	Prob:	Significant/Not Significant
Rural	41.2	43.2	0.51	Not Significant
Urban	42.0	43.5	0.07	Not Significant
Sindh	41.3	43.5	0.01	Highly Significant

Fig 13: Student's Writing Mean Score by Location and Medium of Instruction



8.3 Background Questionnaire

The background questionnaire data from student, teacher and head teacher background data was used. The data items available for these analyses are given below. Further information regarding the independent variables and the regression models used for the analyses can be obtained in the Technical Report and through PEACE.

Student Background Data

- Facilities at Home
- Daily Time for Homework
- Work after school
- Class Teacher
- Separate Language Teacher
- Seating arrangement in the class
- Language Class – Textbook reading

Teacher Background Data

- Gender
- Academic qualification
- Professional qualification
- In service training during the last two years
- Tenure
- Number of years of teaching experience
- Number of years of experience in the current school
- Time for travel to school
- School Gender (Boys/Girls/Mixed)
- Time spent checking homework
- Multi-grade teaching (i.e. simultaneously teaching 2 or more classes)
- Teaching the whole class
- Teaching in small groups
- Did students ask questions?
- Did children read textbook?
- Use of reference books during language class
- Use of textbook during language class
- Use of curriculum during teaching

Head Teacher Background Data

- Academic qualification
- Professional qualification
- Living in the local area
- Time spent to travel to school
- Status of school (Primary/Other)
- School gender (Boys/Girls/Mix)
- Library in the school
- Teacher expectation of students
- Teacher behavior

Below are the most significant findings:

Student Questionnaire Answers

Peaceful place to study at home

It was found that students who do not have a peaceful place to study at home achieve lower scores than those who have a peaceful place to study at home.

Daily time for home work

It was found that students who spent less than 30 minutes doing homework achieved lower scores than those spending more than 120 minutes. Time spent on homework resulted in little significant difference in achievement in the writing tests.

Do you have a class teacher?

The effect of having a class teacher resulted in better scores in reading , while the effect was not significant in writing achievement.

Do you have a separate language teacher?

Having a separate class teacher appeared to have little effect on student achievement and was not significant for the writing achievement of students.

Sitting arrangement in the class – Mat/Rug

Students who were provided with a mat or rug to sit on in the class had better achievement in the Writing – B test only. The effect was not significant for student achievement in the other three tests.

Text book reading in the language class

Student showed better results where the textbook was used often and once a week as compared with once in six months or a year.

Teacher Questionnaire Answers

In service training during the last 2 years

In-service training did not appear to have a significant effect on student achievement.

Years of teaching experience

Students who were taught by teachers with 16 – 20 years of teaching experience achieved lower scores as compared to those being taught by teachers with more than 20 years of teaching experience.

Time for travel to school

It was found that where the teacher was spending between 30 and 60 minutes commuting to school the students achieved higher scores as compared to those students whose teachers were spending more than 60 minutes.

Did you teach a multi-grade group?

Multi-grade teaching (i.e., teachers teaching 2 or more classes simultaneously) had a negative effect on the student achievements. This is a very important finding for making policy decision as 55% of Grade 4 students are affected by multi-grade teaching. This estimate is based on teacher background questionnaire data.

Did you teach in small groups?

Students who were taught in small groups for writing, often or once a week, achieved better scores than those students who were not taught in small groups. Grouping appeared to have little effect on students' reading scores.

Did students ask questions?

The ability of students to ask questions in the classroom appeared to have little effect on student achievement.

Did you use reference books during language class?

Where no reference books were used during the language class student achievement was lower than classes where students used reference books once a week.

Head Teacher Questionnaire Answers

Head teacher academic qualification

Where students were taught by head teachers having a BA or BSc degree the student results were poorer in reading; students who were taught by teachers with an Intermediate qualification had a negative effect on students' writing results.

Do students perform according to teacher expectations?

Students whose teachers had identified that students in their classes performed according to their expectations achieved a better result in the writing test as compared with other students.

9. Constraints

The development in implementing and analyzing a provincial assessment is a complex and challenging task. It involves the training of staff in curriculum analysis and mapping, the development of test frameworks and items, and the organization of the assessment instrument booklets as well as the printing, collating, distribution and collection of the assessment instruments after testing under tight deadlines. The Grade 4 assessment instruments were piloted in November 2009. These were marked and coded by the PEACE team. Data entry was done by the Bureau of Curriculum and item analysis was carried out by the PEACE subject specialists. Between November 2009 and the April 2010 the PEACE team in Jamshoro, with the help of EU technical assistance, reviewed and developed, collated and distributed the assessment instruments and undertook the training of the test administrators and monitors. After the large scale testing in April 2010 the assessment instruments were returned to the district offices where marking and coding took place and the data was sent to PEACE. Data cleaning and statistical analysis was undertaken with the help of EU technical assistance between April 2010 and May 2011.

The **main constraints** identified are listed below:

The Sample

- Difficulties of using the Grade 3 enrolment from the 2008/09 Census of Schools as the measure of size (MOS) for the PPS sampling 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 2008/09 Census of Schools. However, this did not provide accurate information. For example it resulted in some schools which had been identified in 2008/09 Census having 0 enrollment and also schools which were identified as closed, not being included in the school sample despite having an appropriate student population.
- Due to time constraints PEACE was not able to carry out sufficient checking/monitoring of the sample selection.

The Assessment Instruments

- PEACE did not have the services of a full-time In-Page composer for producing the Urdu tests in camera ready form. An In-Page composer is necessary to develop the test items in Urdu. Much of the composing work was done by the subject specialists and by the provision of short term support of a composer from the Bureau of Curriculum.
- The Language test had some misprinting due to the time constraints and the lack of monitoring of the instruments resulting in no corrections being made. This had implications for the statistical analysis of the data.

Statistical Analysis

- PEACE did not have the services of a full-time education psychometrician for analyzing the data. This meant that the subject specialists conducted the analysis through on the job training. This meant that it was time consuming and there were

some delays in the finalization of the analysis. This was the first time the group had undertaken such an analysis from a complex survey.

Logistics

- Time constraint. It was difficult to ensure that all the tasks were completed on time in a rigorous manner;
- The collection of the assessment instruments after testing had taken place was also difficult especially in remote areas;
- 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 PEACE;

Test Administration

- There was little monitoring and accountability undertaken by the districts or PEACE regarding the number of booklets that had been provided to each school and district for testing and the number of booklets that had been returned to the districts or PEACE.
- This time Guidelines to be provided to the focal persons to ensure that they fully understood their role in the test administration

Management

- The monitoring tool was too simple for effective monitoring to take place– it required only a yes or no answer and issues were only rarely identified;
- District focal persons did not always carry out their duties to the full even after briefing by PEACE; guidelines had not been provided for them.

Staffing

- There was not sufficient staff in PEACE to ensure that the assessment was conducted in an efficient and timely manner. Many of the staff members 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 made in connection with the nomination of test administrators and the test administrators were not informed in sufficient time to attend the training;

Storage

- In marking and coding the scripts there was not sufficient space available for storing the assessment materials.

10. Lessons Learnt

- The tightness of the testing programme requires the availability of a full time In-Page and Sindhi composer
- The need for the employment of an independent editor to review the tests before they are printed and also after the first sample printing has been completed in order to ensure that there is no misprinting;
- The need to emphasis that the assessment instruments should be kept securely and that all the assessment instruments, whether used or unused, should be returned to PEACE.
- The need for monitors to be more able to address day to day problems;
- Quality Assurance procedures are required to be put in place for all testing activities;
- The need to disseminate information and results regarding testing to all the stakeholders to enable them to understand the importance of conducting provincial assessments and to encourage them to assist in the activity. Last year's disseminate workshop positively influenced during field testing.

11. Implications

The results of the tests have implications for the development of a quality education in Sindh province.

Strategies and action plans need to be developed to:

- Identify requirements and strategies and plan for improvements in student learning;
- Interpret the National Curriculum according to the needs of the 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.

12. Recommendations

- There is a need for the assessment results to be used to inform policy and teacher's practice in schools. It is recommended that a permanent, effective arrangement, headed by the Director BC&EW with the assistance of PEACE Coordinator, should be established to provide feedback to Districts, teachers for further action to be taken to support teachers. An Additional Secretary (to be appointed) should be responsible for this arrangement.
- Until now, dissemination has consisted of the provision of the reports and leaflet to all district units and officers, a dissemination conference for all districts as well as for

provincial representatives. There is a need for improvements in dissemination beginning with the GoS identifying the purpose of dissemination:

- To improve the knowledge and skill base for making judgments.
 - To educate all stakeholders regarding student achievement.
 - To obtain feedback as to what can be done in response to the information disseminated.
 - To develop cooperation and communication between stakeholders.
 - To make teachers accountable.
 - To make the Directorate of Education accountable.
 - To make **STEDA** PITE, BC&EW, STBB accountable for the quality of their training, curriculum support, textbooks.
 - Dissemination should be conducted to all stakeholders – parents, teachers, districts, the Directorate of Education, the GoS and GoS institutions.
- Reports/sections of the reports should be translated into Sindhi and possibly Urdu.
 - There is a need for the GoS and its institutions to identify and provide support for improvement in schools according to an agreed timetable (Follow-up by improving teaching). This will support the impact of the assessment results and also enable the teachers to identify how they will use the support for improvements to be made. Continuing to disseminate information should also act as a catalyst for improvement.
 - The GoS should provide training to its district officers to disseminate and support improvements in schools.
 - PEACE already provides a Feedback report which identifies some of the teaching improvements that can be made for specific areas. However, this information needs to be used and additional support materials should be provided by PITE, STBB and BC&EW.

DRAFT

ANNEXURES

List of Provincial Working Group Members

Chairman: Secretary Education

Members:	Secretary Education	Chairman
	Director Bureau of Curriculum	Member
	Additional Secretary (A&T) Education	Member
	DG PITE	Member
	Chairman of the Textbook Board	Member
	Deputy Programme Manager, PEACE, RSU	WG Secretariat
	Registrar, University of Karachi	Member
	Coordinator, PEACE	Member
	Deputy Director BC&EW	Member
	Subject Specialists, PEACE	Member*

EU SERTA Quality Education Adviser (Ms L. Jones)
 EU SERTA Statistical Analysis Adviser (G H Choudhry)
 WB Education Specialist (A. Norley)
 WB Education Specialist (Ms Umbreen Arif)
 USAID PRESTEP Education Specialist (Ms Ratan)

*Depending on subject and/or availability

Language Test Specifications

Below are found test specifications of reading and writing.

Table 1: The Number of Items for each Context for Reading and Aspect for Reading Grade 4

Context for Reading	Aspects of Reading									
	Forming a General Understanding		Developing Interpretation		Making Reader-Text Connections		Examining Content & Structure		Total	
	%	No. of Items	%	No. of Items	%	No. of Items	%	No. of Items	%	No. of Items
For Literary Experience	25	7	16	5	8	2	6	2	55	16
For Information	20	5	14	4	7	2	4	1	45	12
To Perform Task	-	-	-	-	-	-	-	-	-	-
Total %	45	12	30	09	15	04	10	03	100	28
Total Items										28*2=56
Grammar										10
Total items										66

Table 2: The Number of Items for each Context for Writing Grade 4

Purposes for Writing	Prompts/Items %	No of Items
Narrative Writing	35%	1
Informative Writing	35%	1
Persuasive Writing	30%	1
TOTAL	100%	3

Test Instrument Development

Item writing for the 2010 Provincial Assessment was conducted in 2009 when Mapping and Item Writing Workshops were held.

The workshop objectives were as follows:

- To develop technical capacity and sustainability for test development in PEACE;
- To map the National Curriculum in the areas of reading and writing;
- To identify and understand the Grade 4 competencies and the hierarchy of abilities;
- To develop Tests Specifications according to the weightage identified in the curriculum documents;
- To develop test items for Language

The participants in the Workshops were members of the Bureau of Curriculum, PITE, Language teachers and PEACE subject specialist.

The workshops were activity based with all participants being involved. The outcome of the workshops was the development of **Language maps, test specifications and items** (350 for each aspect of Language to be tested). These competencies and test items were used for the pilot testing of Language and for large scale testing of Language in 2009.

The **assessment framework** developed was the foundation for the provincial assessment and was the basis for all item development. The language assessment framework is a three dimensional framework. It main concerned is the development of literacy skills (Reading, writing and Knowledge of Language Grammar & Vocabulary).

Reading frame work consists of two main dimensions. Context for reading and aspects of reading

The Content for reading consists of:

- Reading for literacy experience
- Reading for information
- Reading to perform a task

(According to framework “Reading to perform a task” will not be assessed at the Grade-4 level.)

The aspects of reading consist of:

- Forming a general understanding
- **Developing interpretation**
- Making reader text connection

- Examining content & structure

In writing framework three purposes are identified.

1. Narrative
2. Informative
3. Persuasive

To check the area of knowledge of language & vocabulary three aspects of language were used such as vocabulary, Grammar & sentence structure .

On the basis of these domains and the competencies developed, test specifications for Language were prepared and test items developed. The specification table provided a guideline to the development of a comprehensive reliable, valid and practical test for the pilot testing in 2009. The same tests were later reviewed and finalized by Language experts.

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

The development of a test specification ensured that the test measured a representative sample of the curriculum content for the areas to be tested and its objectives. It ensured that the curriculum content was more likely to be assessed in a balanced way. Writing a test specification required:

- ✓ A list of all the instructional objectives and hierarchy. From this specific contexts to be used for testing as well as the aspects to be tested were identified.
- ✓ The content of the subject areas.
- ✓ The weightage to be given to each of the instructional objectives.

From this framework, curriculum competencies and test specification development, **test items** were written. The test items consisted of multiple choice items for the reading tests and constructed response items for the writing tests. Examples of the multiple choice test items used in the tests are found in Annex 7.

Pilot test in Language were developed from the pool of items. Three tests, of 60 test items each for each area to be tested were designed. Items which tested the key competencies in the identified language skills 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.

The tests were further reviewed and printed for pilot testing in 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.

For the piloting, 46 schools were selected from 23 districts (2 schools from each district) A sample of 823 students took part in the pilot testing in 2009 in Language.

From the result of the **statistical analysis** of the pilot items (using ITEMAN software), items were selected and additional items developed to “fill in the gaps”. After item selection and writing additional items, formats for large scale testing were developed and administered in April 2010 to a provincial sample, to establish a baseline for Grade 4 Language achievement.

The items were then organized into four test booklets; RA, RB, WA, WB test booklets were developed so that students would not be able to copy from each other. The items were ordered in each sub-content area, from easy to difficult.

The **Background Questionnaires** for Head Teachers, teachers and students were also developed. These questionnaires focused on such things as school conditions and climate; teachers and teaching practices; Supporting Inputs for Schools; and, students’ home backgrounds. Difficulties were found in constructing some of the questions, as well as in ensuring sufficient coverage of each background and context variable in relation to the length of the questionnaires and the time it would take for the personnel in the sample schools to complete them.

Marking and coding and data entry was completed using manual recording of scores, Excel and SAS software to finalize the data for statistical analysis.

Statistical analysis was completed using the SAS, BILOG_MG and WesVar software.

Some of the **difficulties** experienced in instrument development are listed below:

- Ensuring that there were sufficient items to cover the major part of the curriculum;
- Difficulties in spiraling items so that students in the different booklets were given similar items at the same stages in the booklets;
- The time taken to answer the background questionnaires;
- Difficulties in the distribution of the booklets, Test Administrator Guides and conducting the TA training

DRAFT

Developing the Provincial Test Sample

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 2009/2010 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-Gradeification of District by location type (Rural/Urban) resulting in 46 design strata. At the first stage, schools (or clusters of schools) with Grade 4 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. The measure of size (MOS) for the PPS sampling of schools was the Grade 3 enrolment from the 2008/09 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 2008/09 Census of Schools. Since the 2008/09 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 student during 2009/10 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 2009/10 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-Gradeification 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 3777 schools (PSUs) with 10 students selected per sampled school resulting in an overall sample of about 37770 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 effect². 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-Gradeification 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 2008/09 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 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 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 Madow³ (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 2008/9 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

² 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.

³ Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons

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 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 SAS and the sample of schools was selected using SAS 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 Grade 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 provided to the TA were generated from 11 to 29 and a skip interval was used for enrolments of 30 or above.

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 2009/10 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 2009/10 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 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.

Some of the sampling issues identified were:

- 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 high enrollment were found to have 0 enrollment
 - ✓ Some schools which had been identified in 2008/09 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 recommendations are made for improvement:

- 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.

DRAFT

Annex 5

The Grade 4 Language Analysed Provincial Sample

The following table identifies the proposed number of schools to be sampled, the 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 3603 schools.

District Code	District Name	Area Type	Sampled PSUs	Ineligible PSUs	Respondent PSUs			
					Reading Book-A	Reading Book-B	Writing Book-A	Writing Book-B
Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9
01	Badin	Rural	114	3	109	106	109	106
		Urban	37	0	34	34	34	34
02	Dadu	Rural	140	0	133	132	133	132
		Urban	56	0	52	52	53	53
03	Hyderabad	Rural	27	0	27	27	27	27
		Urban	124	1	120	120	120	120
04	Thatta	Rural	132	0	120	119	116	116
		Urban	30	0	29	30	29	29
05	Mirpurkhas	Rural	97	0	88	88	89	89
		Urban	58	2	54	54	54	54
06	Tharparkar	Rural	125	4	99	94	99	94
		Urban	18	2	16	16	16	16
07	Sanghar	Rural	121	2	115	112	115	114
		Urban	67	0	65	65	65	65
08	Karachi	Rural	17	0	16	16	16	16
		Urban	260	8	259	259	259	259
12	Jacobabad	Rural	92	5	82	81	82	82
		Urban	46	0	40	40	42	42
13	Larkana	Rural	86	0	81	81	81	81
		Urban	101	0	96	96	94	93
14	Shikarpur	Rural	86	5	79	79	79	79
		Urban	57	0	57	57	57	56
15	Khairpur	Rural	155	2	146	145	146	145
		Urban	44	1	43	43	43	43
16	N. Feroz	Rural	131	1	119	118	120	121
		Urban	43	0	38	38	39	39
17	Benazirabad	Rural	107	5	98	94	98	95
		Urban	49	9	39	38	39	38
18	Sukkur	Rural	70	1	64	62	64	62
		Urban	79	2	77	76	76	77
19	Ghotki	Rural	143	0	134	131	133	133
		Urban	42	0	33	32	32	33
20	Umerkot	Rural	108	3	98	96	98	96
		Urban	22	0	19	19	19	19
22	Jamshoro	Rural	49	1	42	42	42	42
		Urban	58	0	54	54	54	54

23	Matiari	Rural	78	2	73	73	73	73
		Urban	39	0	37	37	36	36
24	T. Allahyar	Rural	72	0	67	64	67	64
		Urban	32	0	32	32	32	32
25	T. M. Khan	Rural	64	2	59	58	59	58
		Urban	31	2	25	24	26	25
26	Kashmore	Rural	86	11	71	71	71	71
		Urban	44	1	40	39	40	39
27	Kambar	Rural	104	4	94	94	94	94
		Urban	62	1	61	61	61	61
Sindh			3,603	80	3,334	3,299	3,331	3,307

DRAFT

Coding, Data Entry and Cleaning

For the 2010 Language 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. Marking the writing test scripts was completed according to the writing criteria. 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 Hyderabad. This involved marking and coding of approximately 80 assessment items for each student as well as 13 items in the Head Teacher's Background Questionnaire, 66 items in the teacher questionnaire and 74 items in the students' questionnaire. They were instructed on how to enter the data on the Scoring 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.

Checking the data was an onerous a task so it was not possible to check every single sheet. The data of two students 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 one out of 10 students on each scoring sheet in Hyderabad. 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.

Twelve persons were involved in entering the data in Excel, 12 from the Bureau of Curriculum and GECE colleges. 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.

The Excel 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. After converting these files into SAS files, fractional parts as found in the Excel files were re-coded from A, B, ...F to 1, 2, ...6. Again many errors were found. Finally all SAS data files were combined to make a whole province file for each of the four areas of Language 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:

- 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

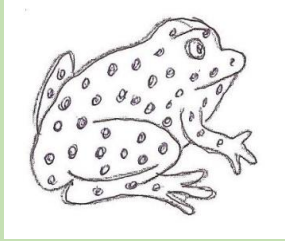
Examples of Test Items

Question 1

آزمائش سنڌي

هدايت:- هيٺ ڏنل ٻنهي عبارتَن کي غور سان پڙهو ۽ ٻئي صفحي تي انهن متعلق ڏنل سوالن جا جواب ڏيو.

عبارت نمبر 2

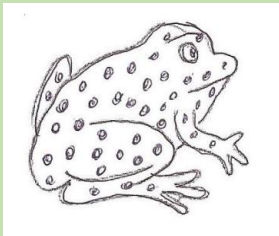


آهي هڪڙو ڏيڏر منهنجو
ننڍڙو تم آ نالو جنهنجو.

تَبَ ۾ کيس جڏهن وڌو
ترندو کيس تڏهن ڏٺو.

ڀاڻي سارو پي ويو آ،
صاحب هڙپ ڪري ويو آ.

تران تران تي تُو زور لڳائي،
نڙيءَ ۾ هڪڙو پلپلو ٺاهي.



عبارت نمبر 1

ڪجهه ڏيڏر آهن.....
تڪرين منجه،
اتي ويهن ٿا.....
۽ سوچين ٿا.

اڪ چنپي دنيا ٿا ڏسن،
ڪڏهن جُهوتن ۾
ڪڏهن جَهپتن ۾
نخرا ڪن ٿا چرڪڙ جا.....
ڪڏهن رانديون لڪَ لڪوڻيون
ڪيڏن.

ٽپا ڏئي ٿا شام جون ڪرن،
گيت سريلا ويٺا ڳائين،
پَلَ ۾ ڏاڍيان.....

پَلَ ۾ ڏيري،
چپ گُهنجائي سڀني ٿا وڃائين،
ڪجهه ڏيڏر آهن تڪرين منجه،
پر.....
سڀ نه آهن!

ٺڪرين منجهه!!!

سوال نمبر 1 نظم نمبر 1 جي مطابق ڏيڏر رهن ٿا۔

الف

درياءَ ۾

ب

ٺڪرين تي

ج

ڍنڍ تي

د

تلاءَ ۾

سوال نمبر 2 نظم نمبر 1 جي مطابق ڏيڏر جهڀڻن ٿا۔

الف

ڪاوڙ ۾

ب

ساهه ڪڍڻ لاءِ

ج

مڪين کي پڪڙڻ لاءِ

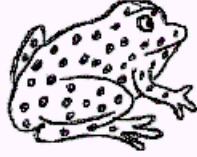
د

خوشيءَ ۾

آزمائش اردو

ہدایت: ذیل میں دی گئی عبارتوں کو غور سے پڑھیے اور اگلے صفحے پر ان کے متعلق دیئے گئے سوالات کے جوابات دیجیے۔

عبارت نمبر ۲



مینڈک میرا چھوٹا سا

تھکانا نام ہے اس کا

شب میں اس کو ڈالوں میں

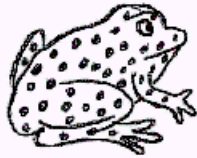
تیرتا پھر تادیکھوں میں

بی لیا اس نے سارا پانی

کھالیا سارا صاف بھی

ٹرانے کو منہ جو کھلا

منہ میں بن گیا ”میلہ بلا“



عبارت نمبر ۱

چٹانوں میں مینڈک بھی رہتے ہیں چند

سکوں کے دریا میں بہتے ہیں چند

خیالوں کی دنیا میں رہتے ہیں گم

پلک ان کو جھپکاتے پاؤں گے گم

جہاں کا نظارہ وہ کرتے ہیں خوب

کبھی مکھیوں پہ جھپٹتے ہیں خوب

کبھی جھوٹا عصہ دکھائیں جناب

چھپائی چھپیں بھی وہ کھلیں نواب

پھریں سارا دن یہ بھدکتے ہوئے

سر شام سوئیں گیت گاتے ہوئے

دا دپائیں یہ گیت گا کر کبھی سے

کبھی چیخ کر کبھی آہستگی سے

لیوں کو سکیریں اور سیٹی ماریں

کبھی اس میں جیتیں کبھی اس میں ہاریں

ہدایت: درج ذیل سوالوں کے نیچے چار چار جوابات دیے گئے ہیں بہترین جواب پر ✓ کا نشان لگائیے۔

سوال نمبر 1- نظم نمبر 1 کے مطابق مینڈک رہتے ہیں۔

دریاؤں میں ا

چٹانوں پر ب

جھیلوں پر ج

تالابوں میں د

سوال نمبر 2- نظم نمبر ایک میں ہے کہ مینڈک چھپتے ہیں۔

غصے میں۔ ا

سانس لینے کے لیے۔ ب

کھیلوں کو پکڑنے کے لیے۔ ج

خوشی میں۔ د

DR

Glossary of Terminology

Assessment	Assessment is the process of documenting, usually in measurable terms, knowledge, skills, attitudes and beliefs of students
Assessment Instruments	These consist of a test framework and specifications, test booklets, rubrics, background questionnaires, guides for test administration, coding, data input and monitoring.
Background Questionnaires	Provide a context for reporting student performance. Student questionnaires collect information on students' demographic characteristics, classroom experience and educational support; teacher questionnaires gather data on teacher training and instructional practices; head teacher questionnaires gather information on school policies and characteristics
Bilog	It models the response of each student of a given ability to each item in the test. Bilog-MG3 converts students' raw scores on a test or versions of a test to a common scale that allows for a numerical comparison between students.
Coding	Each possible answer is given a specific code. For example, 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.
Collapsed School	A school with an MOS of 4, 5 or 6 is collapsed with a neighboring school to form PSUs before sampling takes place
Cognitive Domain	Cognitive domain is knowledge or <i>mind</i> based. It has three practical instructional levels including knowing facts and procedures, using concepts, solving problems and reasoning
Content Domain	Consists of Reading, Writing and knowledge of language
Data Cleaning	The detection and correction or removal of corrupt or inaccurate data from a data set by identifying incomplete, incorrect, inaccurate, irrelevant etc data and replacing, modifying or deleting the dirty data to make a data set that is consistent with other data sets in the task or system.
Data Entry	Data entry is the procedure of transcribing the test data from hard copy to a computer using a specific computer program
Diagnostic	Identifying specific areas of difficulty in students' conceptual knowledge and understanding and their ability to solve problems.
Item	A single question which is written for a specific purpose to measure a specific objective
Multiple Choice Questions (MCQ)	Items which provide a phrase or a stem, and four possible answers from which students select the one correct answer.
National Assessment	Large-scale, sample surveys which assess the performance of the education system and not individual students.
National Curriculum	The curriculum to be followed by all the schools in Pakistan
Not significant	Identifies whether differences in reported scores could have occurred by chance alone, significance tests are reported. A probability where $p < 0.05$ means that the difference could occur by chance alone in 5 out of 100 students. This means that the results are not significant. They cannot provide reliable information for conclusions to be drawn.
Pilot	Where test items and tests are given to a small sample of representative students to see their effectiveness as assessment tools.

Probability	This is the measure of certainty
Provincial Assessment	These are large-scale, sample surveys which assess the performance of students according to specific content and cognitive domains across a range of levels
Primary Sampling Unit (PSU)	A sampling unit created by collapsing small schools and splitting large schools into two or more pseudo schools
p-value	P value is a probability, with a value ranging from zero to one. If, for example, the p value is 0.03, it means that the difference could occur by chance in only 3 out of 100 students.
Questions	Usually a group of test items written to test the same objective
Random Number Table	A table listing random numbers generated by computer software according to specifications provided by PEACE. If the number of children in Grade 4 is 10 or less then all children in Grade 4 are tested; where the number in Grade 4 is found to be 11- 29 the random table is used to select the 10 students to be tested; where the number of students is greater than 30, the skip interval techniques is used to select the 10 students.
Sample	A representative group of students in Sindh Province; representative of each district, rural/urban area, gender (boys/girls)
Significant	Identifies whether differences in reported scores could have occurred by chance alone, significance tests are reported. A probability where $p < 0.05$ means that the difference could occur by chance alone in only 5 out of 100 students; where $p < 0.01$ the difference could occur by chance alone in only 1 out of 100 students (significant difference); where p is 0.000 there is a highly significant difference.
Skip Interval	The ratio of the number of students enrolled in Grade IV divided by the number of students to be sampled from the school, rounded top an integer.
Split Schools	Dividing a large school into a number of pseudo schools by a "conceptual split"
Survey	Same as provincial assessment
Test Framework	Provides the concept behind the testing and details of what will be tested and how it will be tested
Test Specifications	Provides specific information regarding the content and cognitive domains to be tested, item wise
WesVar	A data analysis tool designed for the analysis of a complex survey using replication and regression analysis