EDUCATION MONITOR

Assessment systems in Pakistan: Considerations of quality, effectiveness and use

The Society for the Advancement of Education (SAHE) is a nongovernmental organization established in 1982 by a group of concerned citizens and academics. It builds on the belief that educational justice entails not just access to schools, but to quality education, for all children in Pakistan. SAHE works through an extensive network, the Campaign for Quality Education (CQE), to conduct collaborative research and evidence-based advocacy on key issues to influence educational reform. It has sought such evidence in the realm of data related to school inputs and student outcomes, budgetary analysis, public sector reform and privatization, teacher professional development, language and learning as well as citizenship education.

The report has been produced with the support of Open Society Foundations (OSF). The data and the interpretations in the study are those of SAHE and CQE and do not necessarily reflect the views of OSF.

Published by:
Society for Advancement of Education
65-C Garden Block, New Garden Town, Lahore, Pakistan
www.sahe.org.pk & www.cqe.net.pk

Cover & layout design: O3 Interfaces
EM logo design: Sara Aslam Noorani

EDUCATION MONITOR

Assessment systems in Pakistan: Considerations of quality, effectiveness and use

Society for the Advancement of Education

Campaign for Quality Education

OPEN SOCIETY FOUNDATIONS

FOREWORD

In recent years, assessment has become a buzzword in education reform circles worldwide. In Pakistan too, projects in this area have been implemented since the 1980s, ended and forgotten with little documentation available subsequently. This report documents developments in the field of assessment over the last twenty years implemented by the federal and provincial governments with the support of development partners. Given the technical and multifaceted nature of the present-day assessment enterprise and the role of different players including students, teachers, government and private sector institutions, donors, and political leadership, the authors have covered a vast canvas.

In addition to discussing the historical context in which examination and assessment practices developed in Pakistan, this report gives brief accounts of international practices in assessment design, implementation, analysis, dissemination, use of findings, and impact on learning and teaching practices from Brazil and Uganda. The sections on best practices include the use of assessment findings for curriculum, textbook, and teacher professional development. Most importantly, the report takes a look at how in some countries findings have been used to inform policy.

The report points to the issue of proliferation of the Boards of Intermediate and Secondary Education and, rightly in my opinion, recommends a review of the practice. It also highlights that Pakistan experimented with a national model for sample-based school assessments under National Education Assessment System (NEAS), which was hastily abandoned in view of the imminent passing of the 18th Amendment. Since then, the provinces have been conducting their own sample-based assessments, albeit with varying regularity. It appears from the report that NEAS and the national sample-based assessment have since been revived.

The diverse institutional arrangements, objectives, procedures, and outcomes of the provincial assessments are discussed in the report. Achievements in provincial large-scale assessments since the 18th Amendment seem to be uneven, particularly in sustainable psychometric capacity to design and analyze assessment data beyond averages at gender, grade, and district levels. The report identifies the critical issue of the absence of a career path for technically trained staff in assessment agencies and the detrimental tradition of transfer of such staff to postings where their technical expertise is not utilized.

This report provides a useful introduction to assessment concepts and purposes for practicing teachers, teachers in training, education managers, and administrators at all tiers of government as well as departments of education and psychology in universities. Although students in teacher training institutions and university departments of education and psychology are exposed to courses on testing and measurement, these courses are of little practical value as they are just theoretical and do not mention the actual assessments and assessment practices in Pakistan even in the teaching of validity, reliability, equity, and so on.

This initiative of reviewing the different assessment systems in Pakistan by the Society for the Advancement of Education (SAHE) is both timely and needed. To my knowledge it is a first overview of its kind. I hope it is the beginning of a dialogue on this critical issue of assessments in Pakistan.

Dr. Parween Hasan
Former team leader
National Education Assessment System

ACKNOWLEDGEMENTS

This report was made possible because of the support of many individuals and organizations. The Education Monitor team is grateful to everyone who contributed to this effort.

The publication was made possible due to the generous financial support of the Open Society Foundations (OSF). We are also thankful to all the individuals and stakeholders who participated in this study and whose valuable insights informed the writing of this report.

We would like to particularly acknowledge the support of Ms. Unaeza Alvi, Dr. Fida Hussain Chang, Dr. Nasir Mahmood, Dr. Shehzad Jeeva, Mr. Bakhtiar Ahmad Khattak, Mr. Kamran Lone, Ms. Saima Khalid and Dr. Thomas Christie.

The Education Monitor team

Editorial & Writing: Ayesha Awan, Amal Aslam, Irfan Muzaffar, Abdullah Ali Khan and Abbas Rashid

Research Support: Rafaqat Ali and Lajwanti Kumari

CONTENTS

LIST OF FIGURES, TABLES AND BOXES

List of Figures

List of Tables

List of Boxes

ABBREVIATIONS

ADOE	Assistant District Officer Education
AKU-EB	Aga Khan University- Examination Board
BISE	Board(s) of Intermediate and Secondary Education
BSE	Board of Secondary Education
CCTV	Closed Circuit Television
CTT	Classical Test Theory
DCTE	Directorate of Curriculum and Teacher Education
DFID	Department for International Development (UK)
DLI	Disbursement Linked Indicator
DSD	Directorate of Staff Development
ERQ	Extended Response Question
EU	European Union
IBCC	Inter Board Committee of Chairmen
INEP	National Institute of Educational Studies & Research
IRT	Item Response Theory
KP	Khyber Pakhtunkhwa
NAPE	National Assessment of Progress in Education
NEAS	National Education Assessment System
OMR	Optical Mark Recognition/Reader
PC-1	Planning Commission (Form)-1
PEAC or PEACE	Provincial Education Assessment Center
PEAS	Punjab Education Assessment System
PEC	Punjab Examination Commission
PESRP	Punjab Education Sector Reforms Program
PITE	Provincial Institute for Teacher Education
RSU	Reform Support Unit
SAEB	Sistema Nacional de Avaliação da Educação Básica
SEP	Sindh Education Sector Project (First)
SERP	Sindh Education Reforms Program
SESP	Sindh Education Sector Project (Second)
SLO	Student Learning Outcome
SOP	Standard Operating Procedure
SPE	Supervisor Primary Education
SPR	School Performance Report
SAT	Standardized Achievement Test
Sukkur IBA	Sukkur Institute of Business Administration
TIMSS	Trends in International Mathematics and Science Study

UNEB	Uganda National Examinations Board
UNESCO	United Nations Education, Scientific and Cultural Organization
UNICEF	United Nations Children's Fund
UP	Uttar Pradesh
USAID	United States Agency for International Development

GLOSSARY

assessment: In education, the term refers to the wide variety of methods or tools that educators use to evaluate, measure, and document the academic readiness, learning progress, skill acquisition, or educational needs of students. While assessments are often equated with traditional tests developed by testing companies or institutions and administered to large populations of students, educators use a diverse array of assessment tools and methods to measure students' academic progress. The types of assessments relevant to this report are listed below:

> **formative assessment**: A method teachers use to conduct in-process evaluations of students' learning progress to inform teaching and learning activities.
>
> **high-stakes assessment**: A test used to provide results that have important, direct consequences for examinees, programs, or institutions involved in the testing.
>
> **large-scale assessment**: For the purposes of this report, in the case of Pakistan, the term has been used to refer to the census-like assessments to differentiate them from the set of assessments conducted under NEAS/PEACs which are largely sample-based. However, the term can be used to refer to data collection efforts in which large numbers of students are assessed, through a sample-based or census-based method.
>
> **low-stakes assessment**: A test used to provide results that have only minor or indirect consequences for examinees, programs, or institutions involved in the testing.
>
> **sample–based assessment**: An assessment conducted on a representative portion, selected by an appropriate sampling method, of the target population.
>
> **summative assessment**: A test conducted at the end of a pre-determined instructional period, such as an academic unit or a semester, to evaluate student learning.

assessment framework: A document that defines the purpose of the test and indicates what should be measured, how it should be measured, why it is being measured, and how it should be reported.

classical test theory (CTT): A psychometric theory based on the view that an individual's observed score on a test is the sum of a true score component for the test taker, plus an independent measurement error component. CTT allows for item-level analysis including the difficulty and discrimination of each item, but the item statistics produced by CTT are not independent of the test-takers' characteristics.

cognitive skills: The learning skills — such as one's ability to recall, analyze, and evaluate information —that are seen as crucial to academic progress. These skills are commonly grouped together under the term *cognitive domain*.

comparability: The degree to which two or more versions of a test are considered interchangeable, in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions.

construct: The specific skill or knowledge that the item or test seeks to measure.

constructed response question/item: An exercise for which examinees must create their own responses or products rather than choose a response from an enumerated set. Short answer items require a few words or a number as an answer, whereas extended response items require at least a few sentences.

criterion referenced test: A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparison to cut scores, interpretations based on expectancy tables, and domain-referenced

score interpretations.

cut score: A specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point.

difficulty: *Refer to facility value*

discrimination: Item discrimination is the extent to which test takers with high overall scores get a particular item correct, hence it is the ability of an item to discriminate between low achievers and high achievers. The discrimination index ranges from -1 to 1; however, positive item discrimination is desirable.

distractor: One of the wrong options for an MCQ. An ideal distractor should not be too implausible but should also be indisputably incorrect.

equity: Refer to fairness

equivalence: The process through which two or more test versions are constructed to cover the same explicit content, to conform to the same statistical specifications, and to be administered under identical procedures. Equivalence can be both in the test versions administered in the same year (horizontal) and between years (vertical).

facility value: Facility value indicates difficulty (or easiness) of an item. Index ranges from 0 to 1. Higher indices indicate easier items and lower indices indicate more difficult items.

fairness: As there is no single technical meaning for fairness, brief descriptions of the three most common ways in which the term is used are given below:

> **fairness as lack of bias**: *Refer to item bias*
>
> **fairness as equitable treatment in the testing process**: Fair treatment of all examinees requires consideration not only of the test itself, but also the context and purpose of testing and the manner in which test scores are used. Just treatment includes such factors as appropriate testing conditions and equal opportunity to become familiar with the test format, practice materials, and so forth. In situations where individual or group test results are reported, just treatment also implies that such reporting should be accurate and fully informative.
>
> **fairness as opportunity to learn**: In the context of achievement tests, the test score may accurately reflect what the test taker knows and can do, but low scores may have resulted in part from not having had the opportunity to learn the material tested as well as from having had the opportunity and having failed to learn. When test takers have not had the opportunity to learn the material tested, the policy of using their test scores as a basis for withholding certification, for example, is viewed as unfair.

item: A single part of a test with an individual score; it may be a question, an unfinished sentence, or a single part of a test or questionnaire with an individual score or code.

item attributes: The characteristics of an item, such as its difficulty or discrimination value, that are determined through psychometric processes.

item bias: In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to irrelevant or underrepresented components of test scores that differentially affect the performance of different groups of test takers.

item panel: A small group consisting of three to six people who critically review and refine all aspects of items to ensure that they are of high quality.

item pool: A collection of items tested in a field trial or pretest and of secure items from previous tests that are suitable for use in future tests.

item relevance: The degree to which the knowledge or skill required for answering the item is considered important in the curriculum or to the test taker's real life.

item response theory (IRT): A mathematical model of the relationship between performance on a test item and the test taker's level of performance on a scale of the ability, trait, or proficiency being measured. While item response theory produces item statistics that are independent of the test-taking sample and is regarded as being more applicable to large-scale assessments than classical test theory, its use requires a certain level of skill not widely available in the country.

multiple-choice key: The correct option in a multiple-choice item.

multiple choice questions (MCQs): Items that require students to select the only correct response to a question from a number of options provided.

norm referenced test interpretation: A score interpretation based on a comparison of a test taker's performance to the performance of other people in a specified target population.

objective questions: Test items that require short, precise answers with no room for ambiguity.

optical mark recognition/reader (OMR): A software-enabled process of recording human-marked data into a computer, commonly used in assessment scoring for recording MCQ responses.

pilot test: Another name for a type of trial test that is conducted before the final test, with a small sample of students, to establish the quality and suitability of items, questionnaires, and administration manuals.

proficiency level: An objective definition of a certain level of performance in some domain in terms of a cut score or a range of scores on the score scale of a test measuring proficiency in that domain.

psychometrics: The science concerned with the theory and technique of psychological measurement.

public examination: A type of high-stakes test used for certifying and selecting students, normally held at the end of the academic term or year.

raw score: The unadjusted score on a test, often determined by counting the number of correct answers, but more generally a sum or other combination of item scores.

reliability: The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group.

sample: A selection of a specified number of entities called sampling units (test takers, items, etc.) from a larger specified set of possible entities, called the population.

scale scores: A score to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

score: Points or marks allocated to a student response on the basis of the categories of a scoring guide.

scoring rubric: The established criteria, including rules, principles, and illustrations, used in scoring responses to individual items and clusters of items. The term usually refers to the scoring procedures for assessment tasks that do not provide enumerated responses from which test takers make a choice. Scoring rubrics vary in the degree of judgment entailed, in the number of distinct score levels defined, in the freedom given scorers for assigning intermediate or fractional score values, and in other ways.

standardization: In test administration, standardization refers to maintaining a constant testing environment and conducting the

test according to detailed rules and specifications, so that testing conditions are the same for all test takers. In test development, standardization refers to establishing scoring norms based on the test performance of a representative sample of individuals with which the test is intended to be used.

standardized conditions: Test conditions that are specified in the administration manual and kept the same for all students to whom the test is administered; all students receive the same amount of support, are given the same instructions, and have the same amount of time to do the test.

stem: The part of a multiple-choice item that precedes the options, usually a question, incomplete sentence, or instruction.

stimulus material: Text, diagrams, or charts that provide the context for one or more items.

stratified sampling: A set of random samples, each of a specified size, from several different sets, which are viewed as strata of the population.

subjective questions: Test items that solicit detailed explanatory responses and require a scoring rubric to be marked accurately and fairly.

syndication: A process in which a scorer checks only a single item or a set of items from each paper to ensure that any scoring bias is distributed evenly across all the papers.

test forms: Different versions of a single test seeking to measure the same constructs.

test specification: A detailed description for a test, often called a test blueprint, that specifies the number or proportion of items that assess each content and process/skill area; the format of items, responses, and scoring rubrics and procedures; and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices. *Also known as test blueprint.*

validity: The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test.

Sources: Adapted from Anderson & Morgan (2008); American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (1999); www.edglossary.org

EXECUTIVE SUMMARY

The last few decades have seen a spike in interest in standardized assessments globally. More and more countries have turned towards conducting large-scale student assessments of varying stakes. Unlike ever before, the results of student assessments are expected to inform policy and practice, drive up standards, and fire up the accountability of schools, teachers, and education managers at all levels. Pakistan is no stranger to these trends. Starting with the sample-based assessments conducted by National Education Assessment System (NEAS) in the 2000s and moving on to the large-scale assessments at the provincial level, assessment has become a central focus of education reform.

There are a number of factors which are driving interest in student assessment as a basis for education reform. The first among them, is a demand for comparison of student performance across different groups. Another stimulus is related to the political imperative to implement a uniform curriculum in all schools; thus standardized assessment becomes a tool for ensuring that teachers teach a common curriculum across all schools.

Accountability is yet another noteworthy driver of large-scale assessments. The use of student assessment results as a driver of teachers' accountability is, however, contentious. Some academics think that holding teachers accountable for student performance on the basis of a stand-alone assessment is deeply problematic. Such accountability draws from the assumption that the teacher is the sole determinant of student performance when actually a variety of factors such as the student's socioeconomic background, parental interest, level of nutrition, school environment, and textbooks amongst others may determine performance. The downside of such test-based accountability is that it raises the stakes for teachers and often leads to use of unfair means.

The impact and effectiveness of all interventions that aim at improving the quality of education can be judged on the basis of their impact on learning outcomes. More often than not, monitoring and evaluation plans require the production of student performance data to evaluate existing education interventions. As a result, support for setting up standardized assessments has become a regular feature of education reform projects in Pakistan.

Finally, assessments can play a significant role in informing teaching and learning practices. They can be a driver for improving teaching and learning in the classroom as, after all, what you test is what you get. For instance, improving the quality of the assessments to test higher order thinking skills may actually encourage teachers to emphasize the use of such skills in their classroom practices.

The Education Monitor is an initiative that annually reviews key policies and subsectors of education in Pakistan. This year's Education Monitor explores various aspects of student assessment practices in Pakistan. It traces the emergence of various trends in assessment and provides a comparison of existing practices with best practices to offer insights about potential improvements to assessment systems in Pakistan.

In Pakistan, two very different types of assessments have emerged. There is the system of traditional examinations at the secondary level, which emerged prior to partition. Then there are the sample-based assessments, such as those conducted by NEAS, and large-scale assessments, such as the Punjab Examination Commission (PEC) exam, the Standardized Achievement Test (SAT) in Sindh and the large-scale assessment in Khyber Pakhtunkhwa (KP), at the primary and elementary levels, that are grounded in modern assessment techniques. These emerged when the discourse and practice of standards and standardized assessment began to stream into Pakistan in the wake of global education reforms movements. By the beginning of the 21st century, regular student assessments had become the mainstay of education policy. Both assessment systems are responding to very different guiding principles and purposes. Modern assessments are influenced heavily by global trends of standards and accountability, the results of which are intended to inform policy. Traditional examinations are largely textbook-based high-stakes assessments that mark the completion of the secondary and higher secondary levels of schooling. A bifurcation has occurred between the system of traditional examinations and that of modern standardized assessments. The former continue unabated at the secondary level and the latter are being used at the primary and elementary levels.

The actual practice of assessment has varied tremendously across different countries and often within countries as well. The variation owes largely to variation in the enabling contexts that shape the actual organization of regular assessments. An enabling environment constitutes the extent to which the broader context is supportive of the assessment system. It encompasses political commitment, a strong policy and legislative framework, support of a variety of stakeholders including development partners, favorable institutional arrangements which include a degree of autonomy, stable funding and clear mandate, and competent and permanent

staff.

A review of the assessment systems in Pakistan shows that the degree of political commitment varies. There is a clear commitment amongst the provinces for establishing large-scale primary and elementary level assessments and to use the results of such assessments to improve education service delivery. This focus of provincial governments is driven, amongst other things, by a political desire to implement a common core curriculum to all students regardless of the modality of service delivery. As a result, the need to develop a standardized measure of student achievement has emerged. Development partners' interest is driven by a need to support the government in its own efforts to improve quality education as well as to generate evidence of the effectiveness of their support to the education sector. Secondary level assessments, although well-established, do not receive a similar level of government and donor commitment for reform. The reforms, where they have taken place, are too small and only in the private sector. The Aga Khan University-Examination Board (AKU-EB), which was established to address many of the limitations of the Boards of Intermediate and Secondary Education (BISE) examinations, continues to remain unable to cater to the public sector and no efforts have been made to change this.

Legislation and policy need to go hand and hand. In some cases, one is racing ahead of the other. For example, Punjab has been quick in providing supporting legislation to large-scale assessments. However, the results of the examinations conducted by PEC have not been used to drive improvements in teacher training. In addition to PEC, other institutions, such as the Directorate of Staff Development and Punjab Education Foundation, have been conducting their own assessments in the province. The need for a coordinated assessment policy continues to exist. In KP, there is an emergent policy to use large-scale assessments to drive improvements in the education system. However, KP has no legislation on the large-scale assessment yet. Hence, no institution has the legal mandate to administer the large-scale assessment in KP. Sindh has neither the legislation nor a well-defined assessment policy at present.

In Sindh and KP, the provincial educational assessment agencies, created in tandem with NEAS, are still active. At the national level, NEAS has re-emerged after being dormant for a long time. However, neither the federal government nor the provincial governments have an assessment policy detailing which assessments will be conducted at what level, how results will be used, and for what purposes. Such a policy, once formulated, would help streamline efforts, ensuring greater efficiency in utilization of limited human and financial resources available for assessment activities.

Legislation and policy are necessary but not sufficient conditions for high quality assessments. Human resources matter. Without investing in human resources, assessment practices will not be up to standard. While Pakistan has embraced the promise of modern systems of assessment, it is still catching up when it comes to human resources. This lack of human resource is most pronounced in the BISE which, given their mandate, are not structured or resourced as assessment agencies. Dearth of human resources can be linked to lack of noteworthy professional programs being offered in assessment at institutions of higher education. Opportunities to study educational measurement and evaluation at the university level are limited and are often not relevant to the needs of modern assessments. Assessment agencies may need to determine and communicate their needs to the universities so that relevant high quality programs can be designed. The Higher Education Commission can play a coordinating role in determining the number of assessment professionals needed and call upon the universities to respond to this need.

The quality of examinations within each province will be better assured if the papers are set by a single board of examination instead of multiple boards. In Pakistan, the number of boards has multiplied over time. This increase in the number of boards has been driven by political rather than technical considerations. Therefore, reducing the number of boards may help make better use of scarce human resource available in the country. Ideally, there should be one apex board in each province that is equipped to design examinations according to accepted standards and the remaining boards should only administer the exams.

In order to ensure valid, reliable, and fair assessments, it is critical that the associated assessment practices follow internationally accepted standards and best practice. The assessment cycle begins with the design of the assessment. Key characteristics of good assessment design include validity, reliability, and equity/fairness. An assessment is considered valid when it tests what it has intended to test. It is considered reliable when the scores or results are comparable over time for different test taking populations. It is equitable when it meets requirements of fairness, preventing bias of any form in the assessment design.

The practice of design of good assessment instruments has evolved into a highly refined craft backed by advances in psychometrics (i.e. the science of psychological measurements). A typical design cycle involves contribution from curriculum experts, subject specialists, and psychometricians. Assessment design begins with the development of a test specifications document that specifies the content of the test. Then the items are written, reviewed, and pilot tested for their validity, reliability, and psychometric robustness by teams of re-

viewers and psychometricians. This involves determining the alignment of different items with the curriculum and their difficulty levels. The process must result in an assessment instrument that can reliably distinguish between the abilities of test takers and validly represents the curriculum content that it intended to test.

The design of assessments bifurcates markedly, with more modern professional practices following accepted standards concentrated in the primary and elementary end of the assessment spectrum. The BISE, on the other hand, with the notable exception of AKU-EB, remain firmly ensconced in their decades long tradition of paper setting without recourse to best practices. This is not to say that the rest of the assessment agencies do not need any further improvement. They too remain short of professional human resources needed to adhere to testing standards in letter and spirit.

PEC, Sukkur Institute of Business Administration (IBA), and AKU-EB have clearly developed test specifications, which they have been using for several years now. Provincial Education Assessment Center in KP has also begun to use the standard best practices of assessment design in its sample-based assessments. These agencies train their staff on item development and follow, for the most part, suggested practices for item writing and review. The BISE, however, have a long way to go before transitioning and aligning with established best practices.

Item pilot and psychometric analysis appears to be one component of the assessment design process that requires the most work. Sukkur IBA has been conducting pilot tests since the inception of the SAT; however, it has taken it a few years to improve the rigor and quality of the pilot test. Information on the pilot test sample, analysis, and results have been provided in its technical documentation. PEC has recently begun conducting formal pilot tests; however, details of how it uses pilot findings have not been made publicly available. Similarly, AKU-EB conducts pilot tests but provides limited details on the process and results. Technical documentation is also not available in the case of the large-scale assessment being conducted in KP. For all cases, there is limited information available on the psychometric analysis. While some of these assessment agencies have a good understanding of the standards for assessment design, the lack of publicly available technical documentation means that there is lack of evidence about which of the standards are actually being followed and how.

Assessment implementation is the next stage of the assessment cycle. It refers to administration and scoring of papers. It includes the development and implementation of standard operating procedures for recruiting and training staff to administer and score the test, allocating test centers, distributing and collecting papers, administering tests, and preventing use of unfair means. With regard to scoring of tests, it entails using optical mark recognition software to score multiple choice questions to improve scoring accuracy and prevent unfair means and using detailed scoring rubrics for open-ended questions. Effective implementation requires a quality assurance or monitoring mechanism for ensuring adherence to standard operating procedures and transparency in all practices.

Across the cases, we find slightly more alignment with best practices when it comes to assessment implementation. In most cases assessment agencies make use of existing school teachers to administer and score tests, with the exception of Sukkur IBA who hire alumni to administer the SAT. Using teachers is a common practice in many countries. However, they are usually not practicing teachers or are from a non-participating school. Given the high stakes nature of many of the tests, using practicing teachers whose schools are also being tested can prove problematic as they have a vested interest in the outcomes of the assessment. There is a need to rethink the selection criteria for implementation staff in several assessment systems along with greater monitoring of the implementation process.

Training of the administrative and scoring staff as well as provision of manuals appears to be the norm amongst PEC, Sukkur IBA, and AKU-EB, while the BISE provide no such training for any of their staff. AKU-EB appears to be the only system which conducts extensive training for its administrative staff on how to handle instances of cheating.

The administration of these large-scale assessments and examinations is an arduous task. All assessment agencies appear to have sufficient mechanisms in place for distribution and collection of papers and allocation of test and exam centers. In the case of exams administered by PEC and the BISE, controlling cheating is a major issue due to the high stakes of the exams. While PEC has managed to prevent instances of cheating better, the BISE still struggle. Sukkur IBA faces fewer instances of cheating given the low stakes of the SAT. But, on the other hand, it faces issues related to non-participation. It is clear that AKU-EB has elaborate procedures for dealing with instances of cheating and also far fewer numbers of students to deal with unlike the other BISE.

The test scoring process has traditionally been one of the weaker aspects of assessment systems in Pakistan. This is particularly so in the case of the BISE exams and little has been done to improve these practices. Amongst the primary and elementary level assessments, there appears to be congruence with best practices. Marking of the multiple choice questions using optical mark recognition software has become the norm. PEC's departure from this practice, partic-

ularly given the scale of the exam appears problematic. For marking of constructed response questions, scorers are now provided with rubrics and detailed scoring schemes with the exception of the BISE, which just follow general guidelines. Processes for quality assurance are in place, which entail re-checking a certain percentage of papers.

The final part of the assessment cycle is the analysis and production of assessment results, their communication, and use. Appropriate analysis, meaningful interpretation, and timely dissemination of assessment results are essential for driving improvement in the education system. Assessment results reports must fulfill two conditions. First, they must conform to the assessment framework and second, they must be accessible to a wide range of stakeholders who can potentially benefit from them. Results can be communicated in different ways to different stakeholders. Apart from the main report, assessment agencies produce briefings for ministers or senior policy personnel which focus on key findings and issues along with recommendations; non-technical summary reports that target teachers and the wider population; technical and thematic reports for the research community; and press briefings and media reports.

There are several factors that affect the use of assessment results. First, is the level of integration with the policy process. Legally mandated assessments are more likely to be integrated in the policy process. Other factors also contribute, such as whether the assessment is perceived as a standalone activity or integrated with other educational activities or whether there are actually plans to devise policy or school level actions based on assessment data. The perceived quality of the assessment system is another factor. Lack of confidence in the findings of assessments can be an issue due to the quality of design and implementation of assessments. A third factor, is the effectiveness of the communication strategy. A communication strategy that ensures rapid communication of results, in the form of accessible reports, to all stakeholders is essential.

The priority that is given to the analysis, communication, and use of results varies across assessments in Pakistan. Once again, there is a bifurcation, with the primary and elementary level assessments and AKU-EB placing greater emphasis on result production and communication as opposed to the BISE. There continues, however, to be room for improvement across the different levels of assessment with capacity often lacking in assessment agencies in this critical area.

A review of the cases demonstrates that teaching and learning practices in classrooms, textbook development processes, and on-going professional development of teachers remain largely uninformed by assessment results. There is limited premium on the use of assessment data in determining which aspects of the education system need to be improved and how. As such, there is no robust policy enabling the use of assessment results to drive improvement in the system. In some cases, it also seems that the (perceived) lack of credibility of the assessment hampers use of results.

Challenges exist at both ends. Results need to be produced in a precise and simple manner and communicated to stakeholders in time. At present, assessment results reports take a long time to be produced and disseminated due to lack of capacity within assessment agencies to undertake these activities efficiently. Even when ready, the reports are often inaccessible. With regard to the use of assessments results, planners of teachers' professional development, district managers, and school administrators need to be sensitized, and perhaps even trained on how to make use of the data and results to inform their work, thereby, improving quality in the sector.

Policymakers should make use of data and information on student learning generated from large-scale assessments with great care and caution. There is a tendency to use the results as a proxy for teacher performance, which is not a justifiable policy. Teachers need to be held accountable; however, assessment results are only one source of information amongst many that should be used to measure their performance. It would be best if policymakers deliberated on other means for determining teacher performance and accountability.

To sum up, assessments have a key role in driving quality in the education system. As such, much is to be gained from spending considerable time and effort in improving assessment systems in general. High quality testing is very likely to have a positive effect on the quality of teaching and learning. In order to derive maximum benefit from large-scale assessments, the provinces will need to enforce uniform standards for assessment at all levels of schooling. Sooner or later, the secondary and higher secondary examinations will need to follow the standards and established best practices for design and implementation of assessments and the level of standards for primary and elementary level assessments needs to be raised as well. Given that, the governments should attend to the task of reforming assessments with the urgency it deserves.

01

INTRODUCTION

FOCUS ON ASSESSMENT

The last few decades have seen a spike in interest in standardized assessments globally. More and more countries have turned towards conducting large-scale student assessments of varying stakes. Unlike ever before, the results of student assessments are expected to inform policy and practice, drive up standards, and fire up the accountability of schools, teachers, and education managers at all levels. Pakistan is no stranger to these trends. Starting with the sample-based assessments conducted by National Education Assessment System in the 2000s and moving on to the large-scale assessments at the provincial level, assessment has become a central focus of education reform.

There are a number of factors, which are driving interest in student assessment as a basis for education reform. The first among them is a demand for comparison of student performance across different groups, such as urban versus rural, public versus private, and intra-district comparisons, amongst others. The Annual Status of Education Report and the Learning and Educational Achievements in Punjab Schools study, are examples of reports that rely on comparative statistics based on student scores. The district rankings used in Punjab are also an example of the use of assessment data to justify policy positions based on comparative data. It is felt that assessment results can help inform policy, professional development, and teaching practice in the classroom.

Another stimulus is related to the political imperative to implement a uniform curriculum in all schools. A universal and uniform standardized assessment becomes a tool for ensuring that teachers teach a common curriculum across all schools. Curriculum and standards go hand in hand, and standardized assessments are regarded as the means by which to ensure implementation of curriculum standards.

Accountability is another noteworthy driver of large-scale assessments. Governments elsewhere and in Pakistan have become increasingly interested in test-based accountability. This notion of accountability has come to the fore as part of a larger education reform movement known as the Global Managerial Education Reforms.[1] These reforms use testing along with a potpourri of market and managerialist policy solutions in tandem with ideas such as choice, competition, incentives, and accountability. The use of student assessment results as a driver of teachers' accountability is, however, contentious. Some academics think that holding teachers accountable for student performance on the basis of a standalone assessment is deeply problematic. Such accountability draws from the assumption that the teacher is the sole determinant of student performance when actually a variety of factors such as the student's socioeconomic background, parental interest, nutrition, school environment, textbooks, and a host of other factors may determine performance. The downside of such test-based accountability is that it raises the stakes for teachers and often leads to use of unfair means.

The impact and effectiveness of all interventions that aim at improving the quality of education can be judged on the basis of their impact on learning outcomes. Most education interventions these days have an associated monitoring and evaluation plan. More often than not, monitoring and evaluation plans involve production of student performance data to evaluate existing interventions. As a result, support for setting up standardized assessments has become a regular feature of education reform projects in Pakistan.

Assessments also play a significant role in informing teaching and learning practices. They can be a driver for improving teaching and learning in the classroom as, after all, what you test is what you get. For instance, improving the quality of the assessments to test higher order thinking skills may actually encourage teachers to emphasize the use of such skills in their classroom practices.

The Education Monitor is an initiative that annually reviews key policies, trends, and subsectors of education in Pakistan. This edition of the Education Monitor explores various aspects of student assessment practices in Pakistan. These aspects include looking at the emergence of various trends in assessment and also examining the practices against a set of established best practices. It provides a comparison of existing practices with best practices to offer insights about potential improvements to assessment systems in Pakistan.

The contents of the report draw from extensive reviews of existing research and literature on assessments in Pakistan as well as policy documents and plans, donor project appraisal documents, and data sources. The team also conducted semi-structured interviews with key informants in various assessment agencies and government institutions, international development partners, and other stakeholders.

The second chapter of this report begins by narrating the emergence of various types of assessments in Pakistan. Specifically, it traces the emergence of two very different types of assessments in Pakistan: the system of traditional examinations at the secondary level, and the sample-based

and large-scale assessments at the primary and elementary levels that are grounded in, no matter how imperfectly, modern assessment techniques. Both assessment systems are responding to very different guiding principles and purposes. Modern assessments are influenced heavily by global trends of standards and accountability, the results of which are intended to inform policy. Traditional examinations, on the other hand, are largely textbook-based high-stakes assessments that mark the completion of the secondary and higher secondary levels of schooling.
The third chapter of the report explores the enabling environment of the different assessment systems by way of looking at the factors that support or hinder their development. It uses the examples of Brazil and Uganda to further illustrate the dynamics associated with the establishment of systems of student assessment.

The fourth, fifth, and sixth chapters of the report look at the practices that a typical assessment cycle entail, starting with the design of assessment instruments and then moving onto the conduct and scoring of assessments to the analysis, dissemination, and use of assessment results. The fourth chapter looks at assessment design practices in particular. It highlights the nature of the work required to create valid, reliable, and fair assessments. The fifth chapter looks at assessment implementation practices which include both its administration and scoring. The sixth chapter explores the production of results, their dissemination, and their use by key stakeholders.

The report concludes with recommendations drawn from the discussion in each of the preceding chapters. The Education Monitor team has carefully crafted these recommendations to provide insight about ways to improve the assessment systems in Pakistan. We hope that this effort will help generate debate aimed at improving assessment culture in Pakistan to drive quality education reform.

[1] Global Managerial Education Reforms seek to improve the public sector of education in a country (and thereby, its competitiveness) by instilling new values- ones that promote the adoption of managerial culture from the private sector when approaching reform. Such reforms are focused on 'measurable outcomes, governance and managerial solutions'. These reforms imbibe a standards and accountability approach to improving education quality and are promoted in low-income countries primarily by international organizations and aid agencies (Verger et al, 2013, p.1-3).

02

HISTORY OF ASSESSMENT

INTRODUCTION

Pakistan continued with a traditional approach to examinations when it gained independence in 1947. It was not until the turn of the century that the practice of assessments, sample-based as well as large-scale, streamed into Pakistan's education practices. These assessments, unlike traditional examinations, are grounded in modern psychometrics. These assessments also have a more intricate relationship with education policy than traditional examinations. The results of these modern assessments are intended to be used by a wide variety of stakeholders and not just to rank and sort students. Ideas and practices concerning assessments have reached Pakistan in tandem with the discourse of standards. It is also important to recognize that standardized assessments have proliferated at the primary and elementary levels of Pakistan's education system. This is largely so because of the preferences of donors and development partners who have provided considerable support for primary education over the last two decades. A bifurcation has occurred between the system of traditional examinations and that of modern standardized assessments. The former continue unabated at the secondary level and the latter are being used at the primary and elementary levels. This chapter provides a historical overview of both traditional examinations and modern standardized assessments being conducted in Pakistan.

SECONDARY AND HIGHER SECONDARY LEVEL EXAMINATIONS

EXAMINATIONS PRIOR TO PAKISTAN'S INDEPENDENCE IN 1947

The history of secondary and higher secondary level examinations needs to be understood in the context of the institutions and practices that were in existence in the sub-continent during British colonial rule. The idea of upper school examinations in the sub-continent originated in the mid-nineteenth century British initiatives to introduce such examinations under the auspices of the Universities of Oxford and Cambridge.[1] These universities began school examinations in 1877. Oxford and Cambridge were followed by the University of London, which started a board of school examinations and developed matriculation examinations for university entrance in 1908. These matriculation examinations also served as entry qualifications for local government posts.[2]

In British India, the control and organization of secondary education was initially the preserve of the University of Calcutta. Established on the model of the University of London in 1857, University of Calcutta became the examination and organizing authority for secondary school education in most of India.[3] Around the same time, efforts were also being made to make secondary education independent of university education. In the early years of the 20th century, several successful suggestions were made to make administration and supervision of secondary education independent of the overburdened universities.[4]

In 1919, a commission (which has come to be known as the Sadler Commission after its chairman) was formed by the Indian Government to look into the challenges faced by the University of Calcutta. The Sadler Commission recommended that the line between university and secondary courses should be drawn at the intermediate examination rather than at the matriculation examination and the Government should create a new type of institution called the Intermediate College. The Sadler Commission also recommended the constitution of separate boards for secondary education.[5] The boards of secondary education were created in Uttar Pradesh (UP) and Calcutta on the basis of these recommen-

dations. These boards were statutory autonomous institutions in the public sector. Later, the jurisdiction of these two boards was extended to include intermediate (or higher secondary school) education.

ESTABLISHMENT OF THE BOARDS OF INTERMEDIATE AND SECONDARY EDUCATION

In Pakistan, up until the 1950s, the universities were in charge of intermediate education. A National Commission on Education (which has come to be known as the Sharif Commission) was formed in 1959 by the then president Field Marshal Ayub Khan to review the state of education in Pakistan.[6] Similar recommendations were made to those of the Sadler Commission in 1919. It was after the publication of the Sharif Commission Report in 1959 that intermediate education was placed along with secondary education under the Boards of Intermediate and Secondary Education (BISE). Policy documents at the time highlighted the importance of secondary education as a critical stage in developing people for the workforce and the critical need to improve secondary education and of bringing it up to standard with the rest of the world. These documents emphasized the distinction between secondary education and university education. As the Sharif Commission report put it:

"The first and basic principle is the recognition of secondary education as a complete stage in itself and the need to demarcate it clearly, in respect of objectives, purposes, and methods of teaching curricula and equipment, from university education. It should not be considered to be subsidiary to such education or be designed merely as preparation for it."[7]

The commission recommended that grades 11 and 12 be transferred to the control of the BISE and be regarded as an integral part of secondary schooling. This move was not confined to Pakistan. A similar commission, called the Mudaliar Commission, was formed in 1952 by the Indian Government to make recommendations about reforms in secondary education in India. The Mudaliar Commission also recommended the integration of secondary and intermediate education.[8]

Most of the BISE were set up as a result of the Sharif Commission recommendations in Pakistan. The BISE were not set up to merely serve as testing services but, as the Sharif Commission had recommended, to be responsible for organizing all aspects of secondary and intermediate education. This included curriculum related powers and prescribing courses of study for exams, recognizing and regulating schools including those in the private sector, conducting extracurricular activities as well as retaining functions related to the setting and administration of external examinations at this stage. In practice, however, the BISE have focused almost exclusively on examinations. The examinations conducted by the BISE at the secondary level consist of the Secondary School Certificate for grade 10 and the Higher Secondary School Certificate for grade 12.

PROLIFERATION OF THE BOARDS OF INTERMEDIATE AND SECONDARY EDUCATION

To date, traditional external examinations implemented by the BISE in grades 9 - 12 remain deeply entrenched in Pakistan's assessment system. This is reflected in all of the country's national education policies since independence (see Box 2.1 for further details). Although the policies have, to varying degrees, also called for classroom-based assessments to play some role in evaluating students, they have remained consistent in their support for the examinations implemented by the BISE.

BOX 2.1 THE RELATIONSHIP BETWEEN NATIONAL EDUCATION POLICIES & BISE IN PAKISTAN

Many of Pakistan's past education policies exhibit a tension between preference for a traditional external examination system or a more nuanced internal system of assessment. Yet, notwithstanding the recognitions of their weaknesses, such as emphasis on rote learning, the national education policies have supported the continuation of external examinations administered by the BISE. For example, the report of the first educational conference (1947) observed that the assessment system is:

"Dominated and controlled by externally conducted and impersonal examinations of a single conventional type… in which the teachers and institutions have little or no part". (p.90)

The report called for continuous and periodic classroom-based assessments. This call for internal evaluations and assessments was echoed in the subsequent national education conference held in 1951 that recognized that:

"Few systems of education have been so dominated by external examinations as perhaps that of pre-partition India". (p.35)

However, they also recognized that the main challenge posed by internal evaluations was the lack of standardization across these. The 1957 National Education Policy recommended that public examinations should not be abolished but should be supplemented by a periodic assessment of both school work done throughout the year and behavioral characteristics like honesty and punctuality, initially to be given a weightage of 25%. In 1966, however, the report of the commission on student problems and welfare stated that the practice of taking the marks of internal evaluations into account be discontinued. Instead, it proposed that these internal evaluations be used to determine whether or not a student is prepared to sit for the public examinations. The 1979 National Education Policy also noted the "continuation of an obsolete system of external examinations which encourages learning by rote" (p.66) and recommended that gradually, terminal examinations in grade 10 be replaced by an internal evaluation consisting of periodic tests and a cumulative record of students' aptitude and general behavior.

In a similar vein, the 1998 National Education Policy called for the development of a mechanism in the BISE to report student marks in internal assessments "separately either on the certificates or as a part of a composite assessment" (p.65). The 2009 National Education Policy also called for an appropriate balance between the use of formative assessment approaches and the summative approach of examinations at the higher level.

At present, there are 28 boards in Pakistan in total, including the Federal BISE and the Aga Khan University-Examination Board (AKU-EB). A 29th board is expected to be set up in 2016 in Shaheed Benazirabad in Sindh. These BISE cover both the intermediate and matriculation examinations in the country. Some of the BISE were established as far back as the 1950s and some as recently as 2012.[9] The number of boards has steadily increased over the years (see Box 2.2 for details). These additions have been justified on the basis of an increase in the number of schools and students within the jurisdiction of the existing BISE. It is not clear, however, whether this increase in the number of boards is warranted in light of their mandate to organize secondary and intermediate education. None of their functions necessarily require the development of more than one board of education in each province. Yet, there are multiple boards in each province.

In contrast, states in the Indian Union have continued to work with a single board in each state. In India, a new statutory board is only created when a new state is being formed out of the existing states. Consider, for example, the state of Assam, which was divided into six additional states after independence—Assam, Meghalaya, Nagaland, Manipur, Mizoram, Tripura and Arunachal Pradesh. The jurisdiction of the Assam Board of Secondary Education, which was formed in 1962, had been accordingly reduced with the creation of each new state. On the whole, a single board per state has remained the norm in India. Some Indian states—notably West Bengal, Tamil Nadu, Orissa, Manipur, Kerala, Karnataka, Bihar, and Andhra Pradesh—have two boards, one each for secondary and intermediate education. Also consider the example of the Board of Secondary Education in the state of UP. Established in 1921, this is the oldest board in India and the largest in the world. The population of UP has increased manifold, and is more than the population of Pakistan according to some estimates.[10] Yet, there is only one statutory board of secondary education in UP that caters for the examination needs of the entire population of the state.

There has been steady growth in the BISE in Pakistan between 1950 and 2016. In every decade since the 1950s, new boards have been carved and created across Pakistan. The first board in Punjab was set up in Lahore in 1954.[11] There are now a total of nine boards in Punjab, the newest having been set up in Sahiwal in 2012. Karachi BISE was established in 1950 and later replaced by Karachi Board of Secondary Education and Karachi Board of Intermediate Education in 1974. Boards have been added to the Sindh province throughout the years. Most recently, the Sindh BISE Amendment Act 2015 was passed in January 2016 for the establishment of another board in Shaheed Benazirabad.[12] This will raise the number of boards in Sindh to seven. The first BISE in the then North West Frontier Province (NWFP), was set up in Peshawar in 1961. For almost 30 years, between 1961 and 1990, Peshawar BISE was the sole board in the province. The provincial government constituted seven additional boards after 1990 over a period of 15 years by incrementally reducing the jurisdiction of existing boards. Balochistan currently has three boards. Responsibility for examinations in Balochistan was borne by the Lahore BISE until 1979 when the Quetta BISE was set up. Much later, two more boards were added in Turbat and Zhob.[13] The Balochistan Education Sector Plan 2013 – 2017 had called for the addition of these due to the size of the province and the spread of the population over a large area.

As can be seen in Box 2.2, the greatest proliferation of boards took place in the 1970s. During this decade eight more boards were created across the country. The National Education Policy for this same period (1972 – 1980) had noted that some existing boards are not equipped to deal with the increasing number of students under their jurisdictions and called for the establishment of a board for every 25,000 students with even more boards in areas where the student population was spread over a wide area.[14] This same policy also called for the establishment of a committee comprising of the various chairmen of the boards to ensure "uniformity in standards and procedures".[15] It was followed by the formation of the Inter Board Committee of Chairmen (IBCC) in 1972 as per a resolution of the Ministry of Education.

While the National Education Policy 2009 called attention to

BOX 2.2 GROWTH OF BISE IN PAKISTAN

1950: Karachi BISE set up
1954: Lahore BISE (under the Punjab University Act Amendment Ordinance) set up
1961: Peshawar BISE (under the West Pakistan BISE Peshawar Ordinance 1961) & Hyderabad BISE (under the West Pakistan BISE Hyderabad Ordinance 1961) set up
1968: Sargodha and Multan BISE (under the West Pakistan BISE Multan & Sargodha Ordinance 1968) set up
1973: First Board in Azad Jammu & Kashmir set up in Mirpur
1974: Two separate boards- Karachi Board of Secondary Education and Karachi Board of Intermediate Education (under the Sindh BISE Amendment Act of 1973) set up to replace the Karachi BISE set up in 1950
1975: Federal BISE (under Federal BISE Act 1975) set up
1977 – 78: Rawalpindi and Bahawalpur BISE (under the Punjab BISE Act 1976) set up
1979: Sukkur BISE in Sindh & Quetta BISE, the first board in Balochistan (under Balochistan BISE Ordinance 1977) set up
1982 – 83: Gujranwala BISE (under the Punjab BISE Act 1976) set up
1988: Faisalabad BISE (under the Punjab BISE Act 1976) set up
1989: D.G. Khan BISE (under the Punjab BISE Act 1976) set up
1990: Abbottabad, Bannu, Mardan and Swat BISE (under the NWFP BISE Act 1990) set up
1995: Larkana BISE (under the Sindh BISE Amendment Act 1995) set up
2002: Kohat BISE (under NWFP BISE Act 1990) set up
2003: Malakand BISE (under NWFP BISE Act 1990) & AKU-EB set up
2005: Mirpurkhas BISE set up
2006: D.I. Khan BISE (under NWFP BISE Act 1990) set up
2012: Sahiwal BISE (under the Punjab BISE Act 1976) set up

a reduction in the number of boards to deal in part with the lack of standardization in examinations across the provinces, this recommendation was not heeded, given that additional BISE have been set up since. With the exception of the Khyber Pakhtunkhwa (KP) Education Sector Plan 2010 – 2015, which has noted that the proliferation of boards in KP is one of the reasons for the lack of standardization and quality, any comment on the number of boards is absent in the Punjab and Sindh Education Sector Plans. As discussed above, the Balochistan Education Sector Plan had called for the addition of two boards in the province.

PRIVATE SECTOR PROVISION OF SECONDARY EXAMINATIONS IN PAKISTAN

The only Pakistani private sector provider of examination services at the secondary level is AKU-EB. AKU-EB was formed in 1995 after a consortium of 16 private sector schools requested the president of the Aga Khan University to consider the establishment of an alternative examination board. Aga Khan University's Board of Trustees established a task force in 1998 to assess the viability of an alternative examination service. After carrying out a feasibility study, the task force endorsed the consortium's viewpoint and advised the establishment of such a board. In 2002, the Ministry of Education approved the establishment of AKU-EB as Pakistan's first independent examination board through a government ordinance. The ordinance stated that:

"The Examination Board should be fully autonomous and self-regulatory with the freedom to achieve the objectives for which it is established…. the head of the Examination Board shall be a member of the Inter Board Committee of Chairmen."[16]

After the passage of the ordinance, AKU-EB launched a developmental phase from 2003 - 2007 that involved laying out the organizational structure, recruiting and training staff members, developing syllabi and associated assessment and training material, and seeking out schools interested in affiliation with the newly created board. The implementation of the first three of these processes proceeded largely unimpeded, but affiliation of schools, especially in the public sector, was and continues to appear to be an area of concern. Originally, the ordinance stipulated that AKU-EB may offer its services to all private candidates and non-government schools throughout the country, and schools under the jurisdiction of the federal government were allowed to opt for the examinations as well. The terms of affiliation for schools under the provincial government were not explicitly mentioned, but the ordinance stated that:

"The Examination Board is authorized to expand the provision and scope of its examinations in such manner, by such time, and on such terms and conditions as shall be prescribed by the University and subject to authorization by the Federal Government or the relevant provincial governments for

their respective government schools and institutions."[17]

By 2004, some public-sector schools in Islamabad had registered their interest in seeking affiliation, but before AKU-EB could start its operations, a confluence of external events prevented this from happening. Word spread that elements of AKU-EB's syllabi contained subversions of Islamic beliefs, and several influential voices in the National Assembly and the media sought to discredit the board and demanded its closure before it launched its operations across the country.[18] These voices were further amplified by linking AKU-EB to the distribution of a World Health Organization health questionnaire that contained questions deemed culturally inappropriate for students.[19] Under such a barrage of negative and undeserved publicity, AKU-EB launched a counter-campaign through press briefings, public statements of support from government officials, and assurances that the national curriculum was the only basis used for developing syllabi and assessment material. While these steps were largely effective in assuaging concerns among the public,[20] misconceptions among some schools remained. More crucially, AKU-EB reached an agreement with the Federal Ministry of Education to halt seeking affiliation with government schools and restricted its sphere of influence to the private sector.[21]

AKU-EB conducted its first examinations in 2007, but its outreach and impact remain limited to the private sector due to this limitation in seeking affiliation with public schools.

EMERGENCE OF STANDARDIZED TESTING

Standardized testing has its origins in assessment practices in the west. It has a long history of being used by psychologists for a variety of academic and sorting purposes, but it was only in the 1960s that it started becoming popular as a means to assess students' learning achievements in the United States.[22] It grew in tandem with the discourse of Standards-Based Education. Before it became a global phenomenon, Standards-Based Education erupted in the United States in the 1980s. In 1983, the Reagan administration appointed a commission called the National Commission for Excellence to take stock of the state of education in the country and to make recommendations. The report of this commission, titled 'A Nation at Risk' warned that the United States was fast losing its economic preeminence due to the falling quality of American public schools. It called for a nationwide movement to raise the quality of education in public schools. With the publication of this report the use of standardized high-stakes tests rose steadily.[23] Terms such as 'standards' and 'accountability' entered the lexicon and tests were administered frequently to assess learning in terms of standards being met. The evidence generated by tests pointed out achievement gaps between different racial and socioeconomic groups as a central policy concern, ultimately paving the way for the famous No Child Left Behind Act passed by the Bush administration in 2001. This Act and the resultant state education policies cemented the popularity and use of standardized large-scale testing "galvanizing the assessment movement into a national project."[24] According to some United States-based scholars, gap gazing became the fetish of education research in America and beyond.

While some policymakers supported testing and accountability based on test results, others opposed them. There was a revolt against standardized testing in America in the 1980s with the publication of the Nader Report among others, the major reasons being the "narrow views of ability measured by traditional tests"[25] and that "such tests played a key role in a rigged game, one that favored society's well-positioned elites under the guise of "merit"".[26] Opponents favored performance assessment instead that focused on what the child can do instead of how well he or she can take a test. Standardized tests like the Graduate Record Examinations test and the Scholastic Aptitude Test have been criticized on grounds that scores do not predict academic success and that sub-groups such as women and minority groups tend to perform worse on these.[27] Moreover, they can be harmful to teaching practices and genuine learning by pushing teachers and students in the direction of rote memorization of facts and formulas hindering the implementation of genuine school reforms.[28] However, standardized testing still remains firmly entrenched in the education discourse in the States. As Sacks puts it, "America… is a nation of standardized testing junkies."[29]

Like many other traveling reforms, standardized testing found its way into the global education discourse. However,

as is the case with many such traveling ideas, it was stripped of the debate and controversy that surrounded it in the United States. In the Pakistani context, standardized testing also came to be used interchangeably with any large-scale testing irrespective of whether it had been developed using a standard procedure or not. Standardized testing entered Pakistan's education discourse in the late 1990s and the National Education Assessment System (NEAS) was established in the early 2000s. NEAS was essentially meant to conduct sampled-based and low-stakes assessments. It was assisted in its work by Provincial Education Assessment Centers (PEACs) that went on to conduct independent sample-based assessments at the provincial level. This was followed by the introduction of large-scale testing. Both are discussed in detail below.

At the same time as standardized testing entered the lexicon in the global education discourse, so did an emphasis on primary education in the mid-1990s that kick-started the primary school bandwagon. Governments and civil society in Pakistan too became, and continue to be, focused on primary education. Attention has been turned away from secondary and intermediate education to a large extent even at present.

INTRODUCTION OF SAMPLE-BASED ASSESSMENTS

Sample-based education assessment at the national level

NEAS was set up in 2003 and operated in project mode primarily under funding from the World Bank for five years until 2008 (with a one-year extension till 2009). The idea for NEAS originated in provincial workshops conducted by UNICEF in 1998/1999 on assessment. The National Education Policy 1998 had also called for the development of national assessment capacity and for the monitoring of the aggregate performance of students at the grade 4 level as well as the development of a national achievement test in different subjects at different grade levels.[30] NEAS was also part of the quality assurance component of the Education Sector Reform Action Plan. A concept paper and PC-1 for the establishment of NEAS was developed and approved, initially for 18 months, with funding from the Pakistan Government. The project was then extended with a Department for International Development (DFID) grant through a World Bank trust fund with the Bank providing a loan for this purpose.

Initially NEAS fell under the Curriculum Wing, but was later shifted to the Ministry of Education.[31] NEAS was the central body supported in its work by the PEACs in Punjab, Sindh, then NWFP, and Balochistan as well as Area Education Assessment Centers in Azad Jammu and Kashmir, the Federally Administered Tribal Areas, and the Northern Areas. Cooperation between NEAS and its partners was ensured through the signing of a project agreement in 2003 by the federal and provincial governments and development partners. The World Bank stressed that the Government take ownership of NEAS and its provincial counterparts and all of these with the exception of the Federally Administered Tribal Areas and Sindh were funded as part of the recurrent budget.

The objective of NEAS was to establish an assessment system in the Ministry of Education and to develop capacity to conduct assessments at the national level. NEAS was to monitor whether standards (as laid down in the curriculum) were being met, to work with teachers to enable them to use the data generated from assessments to improve student performance, and to inform policymaking. NEAS conducted four rounds of sample-based assessments in 800 schools in grades 4 and 8 in math, science, language, and social studies during the period 2005 - 2008. During the project period, the World Bank and DFID provided training and technical assistance to build the capacity of NEAS and PEAC staff to design and pilot tests and tools. NEAS was considered to be the only independent quality measure at the time and had considerable support in the donor community that wanted to take its work forward. It was recommended that instead of conducting annual sample-based assessments, NEAS do so every three to four years.[32] However, NEAS became dormant for a few years once the project ended.

Recently NEAS has been revived. It conducted a national achievement test in grades 4 and 8 in 2014 in math, science, Urdu, and English. NEAS is currently preparing for a national sample-based assessment for grade 8 students that it will be conducting in 2016.[33] At present, it falls under the Ministry of Federal Education and Professional Training.[34]

Sample-based education assessment in the provinces

Following the end of the World Bank project in 2008/2009 it was unclear how the relationship between NEAS and the PEACs would evolve moving forward. With the passage of the 18th Amendment in 2010, professional linkages between NEAS and the PEACs became even weaker now that education had become a provincial subject. When formed, most of these PEACs were placed within the provincial Bureaus of Curriculum where they still remain at present.

After the end of the project in 2008, Punjab Education Assessment System (PEAS) received regular funding from the Punjab Government and ran in project mode operating under a PC-1. It received some technical assistance from the World Bank and Program Monitoring and Implementation Unit. PEAS conducted its own sample-based assessment at

the provincial level in 936 schools in Urdu, math, and social studies in grade 4 in 2011. PEAS was part of the Education Department and reported directly to the Education Secretary. PEAS planned to conduct sample-based assessments in other subjects in grades 3 and 4 and envisaged a role for itself in assisting the Punjab Examination Commission (PEC) in exam paper development. However, in 2014, the Chief Minister approved the merger of PEAS and PEC in the province.

PEACE Sindh is housed in the Bureau of Curriculum and Extension Wing in Jamshoro. After 2008, PEACE Sindh developed its own independent diagnostic provincial assessments that it conducted across the districts in 2009 (math test for grade 4), 2010 (language test for grade 4), 2011 (follow-up math test for grade 4, science test for grade 4, and math test for grade 8) and 2012 (follow-up language test for grade 4, social studies test for grade 4, and language test for grade 8). It conducted these with the support of the Education Department and Reform Support Unit (RSU) under the World Bank supported Sindh Education Sector Project (SEP).[35] Regular funding for PEACE Sindh at present comes from the provincial government; it runs in project mode under a PC-1. This funding does not, however, support research and development activities. Donors assist with the training of PEACE Sindh staff, for example, the European Union (EU) previously trained them on data analysis and report writing.

Analysis and reports for the 2009 and 2010 assessments had been completed by PEACE Sindh and disseminated to stakeholders such as the Education Department, RSU, Provincial Institute for Teacher Education (PITE), and the Sindh Textbook Board. However, analysis and reporting for subsequent assessments did not take place due to PEACE's limited capacity.[36] It appears that the World Bank moved away from funding PEACE Sindh because of, "lessons learned from the support of the diagnostic assessment activity under SEP as well as the new priorities that have emerged."[37] The latter referred to support for the Standardized Achievement Test (SAT) under the Sindh Education Reforms Program (SERP) II. This large-scale assessment was prioritized by the Education Department and the RSU over PEACE Sindh diagnostic assessments for future funding and support.[38] There was no support for PEACE Sindh's activities in SERP II. At present, PEACE plays a limited role in the design of the SAT, as an external reviewer of items prepared by the SAT team at Sukkur Institute of Business Administration (IBA).

A taskforce was set up by RSU in 2015 to discuss the need and viability of an independent Sindh Examination Commission that could manage large-scale testing in the province. There was some discussion about such a commission being built out of PEACE Sindh. However, it is unlikely that PEACE will be transformed into the Sindh Examination Commission as it is perceived to be lacking in both capacity and leadership.

PEAC KP is part of the Directorate of Curriculum and Teacher Education. It conducted its first sample-based assessment since 2008 in 2015 where it tested about 4,961 students in grade 2 (who had just entered grade 3) in approximately 350 schools in math, English, and Urdu. PEAC's mandate is still to plan, design and conduct sample-based assessments as well as to build assessment culture in the province. There remains a lack of clarity regarding PEAC's future direction in KP with varying views on the subject. PEAC KP may continue to conduct the sample-based assessment at the level of grade 2 or switch to conducting it in grades 5 and 8 next year. PEAC may also use grade 2 assessment data to identify problem areas and plan timely interventions to address these. PEAC wants to play a part in test design, data analysis, and results reporting of the large-scale assessment (discussed in the next section) being introduced in KP as this capacity is lacking in the BISE (responsible for implementing the large-scale assessment).

PEACE Balochistan is housed in the Bureau of Curriculum. It has been dormant for the last ten years. 2008 was the last year that a sample-based assessment was conducted in the province. The Balochistan Education Sector Plan 2013 – 2017 calls for a revival and institutionalization of PEACE. It calls for the allocation of financial resources to PEACE since it has no funds to undertake further assessments and activities in the province.[39] No such action has been taken yet.

EVOLUTION OF LARGE-SCALE TESTING

The shift towards large-scale assessments in grades 5 and 8 in Punjab, Sindh, and KP has taken place since the early 2000s. While PEC examinations in Punjab have been implemented annually since 2004, Sindh and KP followed suit later in 2012 and 2015, respectively. Large-scale assessments and sample-based assessments are being implemented side by side. Though there was a lull in sample-based assessments by NEAS and PEAC KP, both institutions have experienced a revival of late.

Punjab

The Punjab Government was the first province to organize large-scale assessments at the elementary level. The PEC examinations emerged at a time when the Punjab Government recognized the need for independent data on student achievement as a means for measuring the effectiveness of the system as well as various interventions and inputs. This need emerged partly in response to donor influence.

Prior to this, the Directorate of Public Instruction was responsible for organizing elementary level examinations and

providing guidelines to the district education departments to develop exam papers. However, the lack of consistency and comparability across districts as well as lack of credibility of these examinations was considered to be an issue by many. For a brief period, the Punjab Government discontinued this exam in favor of a continuous assessment policy, which was never implemented. In 2004, the Punjab Government decided on a policy of standardized formal examinations for the purpose of ensuring uniform education standards and providing comparative data on performance of the different districts and feedback to all stakeholders. In 2005, the Directorate of Public Instruction, along with the districts, was charged with the responsibility for executing the exams.

On a parallel trajectory, the World Bank and the Punjab Government signed a credit adjustment agreement supporting the Punjab Education Sector Reforms Program (PESRP). In 2004, the World Bank documents regarded student learning achievement data as crucial to measuring the success of the project.[40] At the time, the World Bank was already supporting sample-based assessments conducted by NEAS and had not foreseen the emergence of PEC. It is likely that this emphasis on achievement data supported the Punjab Government's inclination towards establishing universal standardized grade 5 and 8 exams. Although not part of the original agenda, examination reform came to be seen by the World Bank as an integral part of education quality reforms that included reform in teacher professional development and textbooks.[41] According to them, the data produced by the examinations could be used to establish school-wise and district-wise performance benchmarks.

UNICEF was contracted to provide technical assistance to develop the assessment mechanism. They conducted a review of the examinations along with faculty from Institute of Education and Research at Punjab University, noting that the quality of administration was very good, but the exam papers themselves lacked quality. On the basis of this study, UNICEF recommended the establishment of a central examination authority with specialized personnel to develop the examinations. In 2006, the Punjab Government established PEC to design, conduct, and produce results for large-scale centralized examinations at the level of grades 5 and 8. PEC began regular conduct of examinations, and by 2010 it became a statutory autonomous institution under the PEC Act.

In its early years, PEC received a great deal of technical assistance from UNICEF. This allowed PEC to develop some capacity for designing exam papers and analyzing the exam data.[42] Over the years, PEC began to conduct regular annual examinations and improved its operations. Moving forward, the World Bank, and then DFID, continued to support PEC through the jointly funded Punjab Education Sector Project (PESP) I and PESP II (both these projects provide budgetary support to PESRP).

With time, concerns arose over the quality of the exam papers, administration, and lack of capacity within PEC.[43] In 2011, DFID commissioned an evaluation of PEC design, administration, and dissemination practices. The study found that while administration adhered to Standard Operating Procedures and concerns over cheating were not as rampant as expected, the quality of the examination design, analysis, and dissemination did not adhere to required standards at all.[44] In response, by 2012, the PESP II documents highlighted measures for addressing these issues: market based recruitment of professional staff for test design, piloting of test items, improved procedures and monitoring of test conduct, improved examiner selection, better protocols for scoring questions, and dissemination of results through multiple channels, among others.[45] DFID added a specific Disbursement Linked Indicator to improve the examination and assessment systems as well as to develop an institutional improvement plan for PEC.

Following this, the Punjab Government with donor support brought in Australian Council for Education Research and Cambridge Education to conduct a capacity review of PEC and PEAS and provide recommendations on how to move forward. Numerous recommendations emerged that became part of the institutional improvement plan, some of which pointed to the need for more technical staff, revamping the organizational structure to create clear departments, and improving examination administration processes. One of the recommendations was to merge PEAS with PEC to build up the number of technical staff members. In 2014, this merger was approved by the Punjab Chief Minister. Since then, PEC has continued to work on its institutional improvement plan with technical assistance from development partners.

Sindh

The SAT was initiated in 2011 by RSU and the Education Department in Sindh in order to evaluate student performance in grades 5 and 8 in both public and private schools. Unlike in Punjab, a large-scale assessment was not launched in Sindh with the intention of being a high-stakes test (i.e. linked to promotion and teacher accountability). Rather, their motivation initially was simply to provide credible results and inform parents about the child's true performance, to measure the impact of inputs, and to change teaching practices through the inclusion of items that discouraged rote memorization.[46] Later the SAT was also linked to informing policy makers, "policy decisions would be recommended (on the basis of results) to improve the quality of educational delivery and the learning outcomes of students in public schools."[47]
Prior to the inception of the SAT, there was great dissatis-

faction with the way that school exams were being designed and conducted. School exam results inaccurately showed students performing better than their abilities. These findings were confirmed by a situation analysis of the district examination system at the middle school level conducted between 2006 and 2008 under the United States Agency for International Development supported Links to Learning-Education Support to Pakistan or ED-LINKS project. At the time, only sample-based assessments at the primary and elementary level were being conducted by PEACE Sindh but they lacked credibility in the Sindh Government.[48] It was advised that the Sindh Government conduct a centralized assessment in grades 5 and 8 across the province so that performance of students across districts could be equated and compared.[49]

The SAT was supported by the World Bank and viewed by it as a "monitoring and evaluation activity… to yield relevant, regular, and reliable data to aid the government in evidence-based decision making to improve the quality of service delivery."[50] While officials working at NEAS at the time argued strongly against the need to test every child, low faith in local school exams and equity concerns that all children should receive the same test overrode any opposition. There was disagreement among donors with the EU also initially opposed to a large-scale assessment in Sindh as it felt that sample-based assessments sufficed.[51]

The Minister of Education was especially receptive to the recommendation to launch a large-scale assessment in Sindh. Consultants were engaged to help actualize the assessment; the RSU developed the PC-1 and floated a request for proposals. Although there was some interest in giving the responsibility for this assessment to PEACE, due to its perceived lack of technical capacity, the Sindh Government decided to outsource the entire assessment activity to a third party.[52] Sukkur IBA was selected due to its perceived strengths - experience with designing a university entrance test, familiarity with rural Sindh, and a cost effective proposal. Any concerns regarding its ability to manage the logistics of implementing the SAT were assuaged by the fact that it had plenty of the university's human resources (especially in the form of graduates) at its disposal.[53] In January 2012, a working group for the SAT was set up including officials from RSU, PITE, Bureau of Curriculum and Extension Wing, PEACE, and district education officers from selected districts.

A committee comprising of members of PEACE, the Education Secretary, and some World Bank officials had also been instituted to review the current SAT system to determine "the sustainability and robustness of the testing system, the reliability and validity of test results, and the responsible use and dissemination of test results."[54] Sukkur IBA has now completed four rounds of the SAT. A fifth round of the SAT is expected to take place in 2017 as part of a one-year World Bank project. It is possible, therefore, that Sukkur IBA will be managing another SAT (if its bid is successful).

The RSU also set up a taskforce in September 2015 comprising of persons from the Sukkur IBA SAT team, Bureau of Curriculum and Extension Wing, Sindh Teacher Education Development Authority, PITE, Sindh Textbook Board, Executive District Officers, as well as a number of public and private school directors. The taskforce was to meet on a monthly basis, and establish future direction for the SAT. Two important issues that the taskforce was to determine included who will be conducting the SAT in future, if at all, and, whether the SAT will be made more high-stakes than it is at present.

With regard to who would be conducting the SAT, one possibility under discussion was the establishment of a centralized Sindh Examination Commission to support assessments in Sindh. This notion was supported by the RSU and the Sindh Education Sector Plan. Where to house this commission remained a question. The commission could be a new body distinct from PEACE and the BISE, both of which were considered for the role and would likely be rejected because of lack of credibility.[55] Nonetheless, several stakeholders including RSU, Sukkur IBA, and the EU supported PEACE as they believed it had the requisite technical capacity and felt that further capacity could be built to take over the SAT. Others felt that there was now no need to conduct another SAT. Given that all rounds of the SAT indicated abysmally low student performance, another SAT was not needed to confirm what stakeholders already knew.[56]

The taskforce has decided to set up an independent Sindh Examination Commission. However, the taskforce has not elaborated how this commission will be set up and when. Over the last few months, the taskforce has been dormant and has not met. The taskforce is now likely to meet in May 2016. However, clarifying its plans for the establishment of the commission is not on this meeting's agenda.[57]

Khyber Pakhtunkhwa

The KP Education Sector Plan 2010 - 2015 called for the introduction of a centralized examination system in grades 5 and 8. It noted that students in the province appear for their first external examinations in grade 9. Until then, they are only exposed to internal cluster level exams that are developed by teachers, that are poor in quality and do not encourage development of higher order thinking skills. Due to these concerns, and in a similar manner as other provinces, the KP Government decided to embark on developing a system of centralized examinations.

At present, the system and structure for this large-scale assessment is taking shape. Large-scale assessments were conducted in 2015 and 2016 in KP but only in grade 5. These have been trial rounds to refine the assessment processes. In 2015, due to the perceived lack of capacity amongst government institutions (i.e. PEAC and the BISE), the assessment was outsourced to a private Islamabad-based organization, National Education Evaluation Foundation. However, due to issues with exam design and conduct, the results were not considered credible and, therefore, were not officially announced or shared. After this experience, and given that the cost of outsourcing this assessment would be too high (it would cost approximately PKR 230 million), the Education Department decided that in 2016, the BISE would take charge of this large-scale assessment. Therefore, in 2016 the BISE has been responsible for administering the assessment to grade 5 students across the province with the assistance of PEAC in test design.

From 2017 onwards, the large-scale assessment will be administered in both grades 5 and 8 across the province and will be a high-stakes examination (i.e. promotion to the next grade will be tied to performance). Cluster-based examinations will, therefore, be phased out in 2017. It is expected that the Peshawar BISE will be the main assessment agency responsible for all aspects of the examination process, that is, design, administration, scoring, analysis, and dissemination of results. Capacity building of the BISE is planned under the KP Education Sector Program so that it can take on these responsibilities. However, the BISE are not in favor of being responsible for large-scaling testing of students in grades 5 and 8 and are concerned over the increase in workload.[58]

CONCLUSION

This chapter provided a historical overview of the assessments inherited at the time of partition, which comprised exclusively of external board examinations at the secondary and higher secondary levels, and the developments since. While the traditional examinations continue to exist in present-day Pakistan, the assessment system has evolved, especially since the early 2000s, to now include other forms of assessment at the primary and elementary levels. This chapter traced the rise of sample-based (NEAS and PEAC) and large-scale standardized assessments in Pakistan that varied from low-stake tests (e.g. the SAT in Sindh) to high-stake examinations (e.g. the PEC exam). The discussion in this chapter pointed out that the evolution has, however, primarily been the product of developments in thinking and practices on assessment happening elsewhere around the globe, particularly in the United States. The focus on standardized testing in Pakistan has not grown organically. Rather, it has permeated local discourse through traveling reforms introduced and supported by development partners. The assessment landscape continues to evolve, particularly, with the revival of sample-based assessments conducted by NEAS and PEAC in KP, the institutionalization of the SAT in Sindh, and the introduction of the large-scale assessment in KP. While some assessment systems are now firmly entrenched in Pakistan, such as those of the BISE and the PEC which have legislative cover, other assessment activities are taking place in project mode.

[1] Riding & Butterfield (1990)
[2] Basu (1867)
[3] Ibid.
[4] Sharma (1991)
[5] Sadler (1919)
[6] Ministry of Education (1959)
[7] Ministry of Education (1959), p.113
[8] Ministry of Education (1953)
[9] "Board of Intermediate and Secondary Education Sahiwal" (n.d.)
[10] Population of Pakistan is 180.44 million while that of Uttar Pradesh is 199.6 million (http://www.worldpopulation.com and http://upgov.nic.in/upstateglance.aspx respectively)
[11] The Lahore BISE became responsible for secondary and intermediate examinations for the regions of Punjab, Baluchistan, Azad Jammu and Kashmir, northern areas and overseas candidates. Previously these examinations fell under the ambit of the University of Punjab. At present the jurisdiction of the Lahore BISE is only districts Lahore, Kasur, Sheikhupura and Nankana Sahib
[12] "New BISE on the cards" (2015)
[13] "Education Boards in Zhob" (2011)
[14] The Education Policy 1972-1980 (1972)
[15] The Education Policy 1972-1980 (1972), p.32
[16] Ordinance no. CXIV (2002)
[17] Ibid.
[18] AKU-EB former staff interview (11 March 2016)
[19] USAID (2008); Ali (2012)
[20] USAID (2008)
[21] AKU-EB former staff interview (11 March 2016); USAID (2008)
[22] Sacks (1997)
[23] Moses & Nanna (2007)
[24] Moses & Nanna (2007), p.56
[25] Ibid.
[26] Sacks (1997), p.26
[27] Young & Kobrin (2001)
[28] Sacks (1997)
[29] Sacks (1997), p.25
[30] National Education Policy 1998-2010 (1998)
[31] Tirmazi (2008)
[32] Ibid.
[33] "NEAS Activities" (n.d.)
[34] "About NEAS" (n.d.)
[35] According to World Bank (2013), SEP lasted from June 2009 – June 2012. The World Bank provided approx. US$ 300 million under SEP. As part of the agreement, baseline measurements for diagnostic mathematics, language, science and social studies would be established for class 4 students in a district representative sample of public schools in 2008/9 2009/10 2010/11 and 2011/2012 respectively.
[36] World Bank (2012d)
[37] World Bank (2012c), p.42
[38] World Bank (2012a); World Bank (2012d)
[39] PPIU, Education Department, Government of Balochistan (2013)
[40] World Bank (2004)
[41] World Bank (2007)
[42] World Bank (2006)
[43] World Bank (2012b)
[44] SAHE (2011)
[45] World Bank (2012c)
[46] RSU (2011)
[47] RSU & Sukkur IBA (2012), p.4

[48] According to World Bank (2013), "Critically, government ownership of and commitment to the assessment activity undertaken by PEACE has remained weak. Furthermore, unless overcome, binding constraints arising from the standard workings of the government make a specialized and skilled activity such as testing difficult to sustainably run wholly within a government entity."(p.34)
[49] former RSU staff interview (1 December 2015)
[50] World Bank (2013) p.16
[51] former NEAS staff interview (30 November 2015)
[52] former RSU staff interview (1 December 2015)
[53] RSU staff interview (30 November 2015)
[54] World Bank (2013), p.7
[55] former RSU staff interview (1 December 2015)
[56] Sukkur IBA staff interview (9 May 2016)
[57] Sukkur IBA staff interview (2 December 2015)
[58] Peshawar BISE staff interview (26 February 2016)

03

ENABLING CONTEXT FOR ASSESSMENTS

INTRODUCTION

Quality of educational processes and outcomes cannot be ascertained without assessment of student learning gains. International policies such as Education for All and Millennium Development Goals have also prompted countries to focus more on quality, and, with that, on the need to assess student learning. As mentioned earlier in Chapter 2, the standards movement has also provided impetus for assessment reforms. However, the actual practice of assessment has varied tremendously across different countries and often within countries as well. The variation owes largely to variation in the enabling contexts that shape the actual organization of regular assessments.

The enabling context is one of the key drivers of quality in assessment systems. An enabling environment constitutes the extent to which the broader context is supportive of the assessment system.[1] It encompasses political commitment, a strong policy and legislative framework, support of a variety of stakeholders including development partners, favorable institutional arrangements which include a degree of autonomy, stable funding and clear mandate, and competent and permanent staff. This chapter will provide further details on each of these aspects using the examples of Brazil and Uganda. It then moves on to exploring the enabling context of the various assessments being conducted across the provinces of Pakistan.

ENABLING FACTORS

Political commitment, policy and legislation

In many countries, strong political commitment for educational reform backed by sound policy and legislation is the first step in creating an enabling environment for an assessment system. Often an assessment policy has come about as part of a larger education reform movement within the country or in response to international commitments.

An assessment system is usually backed by law, which authorizes the agencies to organize assessments. Existence of laws also makes the systems stable and permanent as in the case of the Dominican Republic and Peru.[2] In fact, even if there are unstable political conditions, if the assessment agency is mandated by law, it continues to perform its function. For instance, the political instability in the Dominican Republic has not affected the implementation of national tests due to the existence of the law supporting these tests.

The cases of Brazil and Uganda are instructive in the way they have successfully implemented major reforms to their assessment systems and have institutionalized assessments, the results of which are widely used and appreciated amongst stakeholders (see Box 3.1). In Brazil, there was the political will to make education a priority and transform its evaluation; this resulted in new policies and legislation. Revamping the assessment system came as part of a larger education reform effort in Brazil. Legislation was put in place making education and the need to monitor education quality a national priority. This legislation also made the federal government responsible for evaluating education quality. Political leadership at all levels, especially from Brazil's president at the time, along with collaboration with the key assessment agency, National Institute of Educational Studies and Research (INEP), and between national, state, and municipal governments allowed this transformation by providing a supportive environment for assessment activities.

BOX 3.1 HISTORY AND CONTEXT OF ASSESSMENTS IN BRAZIL AND UGANDA

Brazil

Brazil went from having a weak assessment system in the 1980s to a robust system by the 2000s. It began by transforming its federal assessment agency, the National Institute of Educational Studies and Research (INEP). In 1998, Brazil introduced the National University Entrance Exam, one unified exam which selects students for university and certifies completion of secondary education. It was backed by law, but universities could use it voluntarily. Initially universities were slow to adopt it, but by 2009 it was widely used and accepted by universities and students.

In 1995, Brazil introduced national large-scale assessments (refer to large-scale assessments in the glossary), Sistema Nacional de Avaliação da Educação Básica (SAEB), with the purpose of monitoring education quality. Pilot assessments began sporadically prior to this, but in 1995, the Ministry of Education took control of the assessment and ensured greater technical sophistication in the selection of samples, item writing, and analysis. In 1997, the responsibility for the SAEB was transferred to INEP. The SAEB is administered in grades 5, 9, and 12 in all schools, measur-

ing student performance in core subjects of math and Portuguese regularly as well as other subjects from time to time.

Due to demand for school level results for monitoring and accountability, Brazil introduced Prova Brazil in 2005, a census-based assessment administered every 2 years in core subjects in grades 5 and 9 using the SAEB tools. Data from this assessment is used to create a school quality index that combines student performance and repetition rates for greater school accountability and monetary incentives. Both SAEB and Prova Brazil have contributed towards raising awareness of student performance issues.

Finally, Brazil has also created a school-based assessment program, Provinha, introduced in 2008. INEP designs these assessments and schools administer them as a diagnostic at the beginning of the year and at the end of the year to measure whether learning goals have been met. However, its benefits are not clear as stakeholders have mixed views about the usefulness of these assessments.

Uganda

In Uganda, we find a firmly established tradition of examinations inherited from its colonial days. Since 1983, annual exams at the end of each cycle - elementary (grade 7), lower secondary (grade 11), and higher secondary (grade 13) - were made the responsibility of the Uganda National Examinations Board (UNEB). These exams are used for student certification, tracking, and selection in higher education. There are issues with the exams, such as malpractices, teaching to the test, and narrowing of the curriculum due to high stakes. However, over the years reforms have been implemented to mitigate these issues, such as changing the nature of questions to test higher order thinking skills, and adding scores from classroom assessment in the overall grade.

National large-scale assessments emerged in Uganda in the 1990s due to the need for a mechanism for monitoring education quality. Prior to this, it had used examination data for such monitoring, but that had its limitations as the exams were primarily meant for certification. Currently, the National Assessment of Progress in Education (NAPE) is a sample-based assessment conducted in grades 3, 6, and 9. Stakeholders have found the data produced by this assessment to be useful.

Sources: de Castro (2012); Kanjee & Acana (2013)

In the case of Uganda, several factors have enabled assessment reform. There was leadership and political commitment at the highest levels such as the Ministry of Education and Sports, to reform assessment. Uganda's commitment to global policies, such as Education for All and Millennium Development Goals, highlighted the importance of assessment data. Development of the assessment system was also supported by a strong policy framework in the form of the Government White Paper on Education of 1992 which called for addressing inadequacies in the assessment system. This paper along with associated laws allowed for the creation of the national large-scale assessment program.

Support of development partners

In many instances, development partners also appear to play a key role in supporting the establishment or revamping of student assessment systems. Often, countries do not have the capacity or necessary resources to begin large-scale assessment initiatives. Development partners' support has played a substantial role in the initial setting up of institutions and mechanisms for conduct of assessment activities.

Donor support was a critical factor in the reform of Brazil's and Uganda's assessment systems. In the 1990s, the United Nations Development Programme and the World Bank supported assessment activities through funding and technical assistance. The World Bank funded the SAEB assessment in Brazil from 1995 to 2000. Gradually, the Ministry of Education took on more responsibility and eventually took over the conduct of assessments in its entirety. In the case of Uganda, development partners such as UNICEF, UNESCO, and the World Bank highlighted the importance of assessment. Along with this, the World Bank provided funding for the first three rounds of the National Assessment of Progress in Education (NAPE) and technical training to Uganda National Examinations Board (UNEB) staff.

Institutional arrangements

Institutional arrangements vary widely from country to country. The assessment agency is the ministry of education, an independent or semi-autonomous organization, a university, or a private testing services provider. Research studies find that autonomous assessment agencies are usually more stable, functional, and have technical legitimacy.[3] The risk with external agencies is a potential loss of useful communication between the assessment service providers and the ministries of education, thus, reducing the desirable impact of the results of assessment on policymaking and implementation. Different countries have taken different paths to developing successful assessment agencies. In several cases, assessment agencies are located outside the ministry of education such as the INEP in Brazil and UNEB in Uganda. However, in the case of Chile, although the assessment agency was initially independent, it was eventually subsumed under the ministry. This did not, however, limit its functionality as the agency had already developed a great deal of legitimacy.

A clear organizational mandate backed by law is an essential

part of the enabling environment. The mandate includes the types of assessments, their purposes, target population, and frequency. Assessment systems can have a variety of purposes such as certifying completion of a level of schooling and promotion to the next level, identifying gaps in Student Learning Outcomes (SLOs), and evaluating the effectiveness of interventions in the education sector. It is also important that the assessment agency has adequate funds to carry out its mandate.

Stable and substantial funding is a critical factor in developing sound assessment systems. Financing of assessment can be an issue. Many of the assessment systems in Latin America, such as Ecuador, set up with donor funding in project mode and technical assistance, were unable to sustain when donors pulled out.[4] In order for assessment systems to sustain, governments often plan for financing assessments after support from development partners has ended.

In the case of Brazil, transforming the federal semi-autonomous agency, INEP, was a vital step in driving assessment reform. The INEP, which had existed for more than 50 years, was given a new legal mandate in 1997. It clearly outlined the purposes, target populations, and frequency of different assessments. Over the years, INEP's mandate was expanded to manage several of the key assessments in Brazil. Expansion in its mandate also accompanied an increase in its funding- INEP's budget quadrupled in a decade. These steps reduced Brazil's reliance on international partners.

The UNEB is one central assessment institution in Uganda, which has allowed for great efficiency, cross fertilization, and synergy between different types of assessment. UNEB was established through the passage of a law in 1983. Starting with just end-of-level exams, UNEB has slowly become responsible for national and international large-scale assessment programs. It is highly institutionalized and has clear governance. UNEB income comes largely from examination fees, and to some extent from government grants and donor funds. Although the World Bank provided initial funding, the Ministry of Education has been providing funding since then.

Institutional structure and human resources

Well-functioning assessment agencies around the world are highly structured organizations with clearly defined functions for each department. The departments are typically led by appropriately qualified and experienced individuals. Although there are variations in what aspects of the assessment process these agencies handle directly, in-house technical staff usually manages design and analysis and there are affiliated departments or sections that deal with these processes. For example, there are departments for psychometrics, test development, analysis, and research.

Effective assessment agencies have trained professionals to manage and guide test construction and analysis processes. Selection of key technical staff is done on the basis of technical skills rather than simply seniority. Such organizations often have a stable set of in-house technical staff and a combination of subject specialists, psychometricians, data analysts, and statisticians. Continuous professional development is considered essential, provided through on-the-job training and short courses. Many successful agencies offer competitive salaries and have well-established career paths for technical staff. These serve as incentives for retaining such staff in the long run.

Certain staff can be temporary hires such as item writers or scorers. Item writers are sometimes practicing teachers. They are usually trained in how to analyze learning objectives in the curriculum, write items that provide adequate information, and judge the quality of pilot test items in terms of both content and statistical properties. Scorers often have adequate professional expertise and receive training in scoring the particular assessment.

In the cases of Brazil and Uganda, both institutions have well-developed organizational structures that enable them to carry out their responsibilities. A key step in revamping both the INEP and UNEB was strengthening their human resource. Initially, between 1997 and 2006, INEP hired temporary staff in cooperation with United Nations Development Programme and UNESCO. However, in 2007, a specific career path was developed for civil servants in the INEP and emphasis was placed on hiring assessment and curriculum experts rather than just public administrators. INEP now has over 300 permanent staff members with established career paths. INEP is now well-regarded in Latin America for its high quality work and capacity to design and implement a variety of assessments.

Similarly in Uganda, UNEB has stable qualified staff, which is possible because of the status accorded to UNEB, and competitive salaries allotted. Staff also receives many opportunities to learn from participating in international assessment, and opportunities for professional development within the organization exist.

Assessment perceptions and result use

The perceptions of assessments as well as use of assessment results to inform policy, pedagogy, and to some extent, accountability, play an important role in building an assessment culture which, in turn, play a role in the enabling context.

In the case of Brazil, initially the perception of large-scale assessments was quite negative. This changed as the assessments gained legitimacy. The dissemination of school quality

indicators based on assessment results encouraged a demand for such information. The results are now widely used and discussed; the National Congress organizes public hearings to review results, mayors and governors are concerned with results, the media regularly covers this issue, even the teachers unions who were opposed to large-scale assessment now recognize their legitimacy.

In Uganda, NAPE data has been used to monitor and identify trends in student performance and design interventions to raise the standards of student learning. Such demonstration of assessment data use has highlighted the importance of such assessment and contributed towards its institutionalization.

PRIMARY AND ELEMENTARY LEVEL ASSESSMENTS

PUNJAB

Political commitment, policy and legislation

In the case of Punjab, political commitment for assessment reform exists. The establishment of the Punjab Examination Commission (PEC) was part of a wider movement to improve quality of education in Punjab. In 2003, the Punjab Government launched the Punjab Education Sector Reforms Program (PESRP), one of the three pillars of which was quality of education. The Punjab Government has consistently pursued this reform agenda, despite changes in government, and is now implementing PERSP II with the support of development partners. More recently, in 2011, the Chief Minister of Punjab launched the School Reforms Roadmap, which has provided political commitment and leadership for education reform and driven the pace of these reforms through its monthly stocktake meetings. The recent wave of efforts to strengthen PEC has also come in part from the leadership provided by the Roadmap. Ideally, PEC exam data should be used as a means for measuring reform impact, but credible data is required for this. Hence, improving the capacity of PEC and its design and implementation processes has become a priority.

The Punjab Government established PEC in 2006 through a notification and the Punjab Provincial Assembly accorded it legislative cover after the passage of the PEC Act in 2010. The Act governs the composition of the Commission, its functions, delegation, budgeting, auditing along with general rules and regulations. This legislation has ensured the stability of PEC exams, enabling PEC to conduct annual exams consistently over the years.

Although there is a reform agenda backed by donor support and strong legislation, Punjab still lacks a clear assessment policy for the province. There are several assessment efforts in the Punjab conducted in the public and private sectors, at different grade levels, and often for overlapping populations at the primary and elementary levels. According to Department for International Development (DFID), "the existing efforts appear to be duplicative, are not coordinated, and do not form part of a coherent strategy or approach to student assessment in Punjab. There is already a recognized need for clarity of purpose, coordination, and vision to join these efforts up and increase efficiency and effectiveness".[5] A clear assessment policy would help address this issue.

Support of development partners

Development partners have supported the grade 5 and 8 exams in Punjab from their conceptualization to the establishment and strengthening of PEC. UNICEF supported the establishment of PEC and provided substantial technical assistance in setting up the institution, developing practices and procedures. The World Bank has also supported the Punjab Government's reform agenda (PESRP I and PESRP II) from 2004 to date. Since 2012, DFID has also supported the reform agenda as well as the Chief Minister's Roadmap. Both the World Bank and DFID have provided significant budgetary and technical support to PEC. An example of the latter is bringing in organizations such as the Australian Council for Education Research for reviewing PEC's effectiveness. The donors have also supported the development and implementation of an institutional improvement plan, which includes improvement of all assessment practices (details have been provided in Chapters 4 - 6).

Institutional arrangements

Punjab chose to establish a new autonomous institution to run the PEC examinations. This choice provided room to establish an entirely new system for these examinations. There is a fair amount of oversight in the affairs of PEC by the Punjab Government, which plays a significant role by appointing the Chief Executive Officer and nominating the Chairperson and all members of choice (i.e. those whose designations have not already been specified) to the Board of Governors. Yet, PEC remains independent of the School Education Department.

PEC is fully funded by the Punjab Government. It has had stable funding since inception. With a budget of PKR 908 million in 2015 - 2016, funding has increased 15% from the previous year and 65% overall from 2010 - 2011.[6] It is likely, that the increase in budget has been to implement the improvement plan which has, among other things, meant an increase in staff and salaries in recent years. Whether this budget is sufficient to fulfill PEC's mandate is unclear.

PEC has a broad mission of which its key functions include: (1) designing, implementing, monitoring, and evaluating a system of examination for elementary education; (2) formulating policy and programs for the conduct of elementary examinations; (3) collecting data for research to improve curricula and teaching methodology; (4) identifying areas where improvement in the training of teachers or educationists is required; (5) promoting public discussions on issues pertaining to elementary education; and (6) advising the Punjab Government on all policy matters relating to the objectives of the Commission. In practice, PEC's primary function has remained to design, conduct, and produce results of the examinations. PEC has primarily produced results to inform students, schools, and district education departments. In recent years, PEC has begun to provide SLO-based results which may inform teacher training and practice. However, on the whole PEC does not fulfill its mandate. Informing policy and practices in the education sector, delving further into the data for research purposes, and promoting public discussion on issues related to primary and elementary education have, thus far, remained outside the scope of PEC's activities. It is likely this has to do with a lack of priorities amongst the leadership of PEC and the institutions that govern and support it. Thus, adequate staff has not been hired to fulfill these aspects of the mandate.

Institutional structure and human resource

From its inception, PEC has relied heavily on technical assistance from donors for all its key functions including exam paper construction, analysis, and reporting. In its early years, PEC received substantial support from UNICEF in the form of two in-house consultants. More recently, under the World Bank and DFID supported PESP projects, PEC has continued to receive technical assistance in improving their key processes.

Up until 2014, PEC had departments for operations (test administration), research (test construction and analysis), and information technology (data management and producing results). There was lack of clear separation of roles and responsibilities amongst staff and departments, particularly, test construction and analysis. Due to lack of staff, at the time, there was only one research officer responsible for both activities. In 2014, PEC revamped its organizational structure and created more posts for technical work. It also separated the post of director of assessment development from the post of director for research. It increased the number of assessment experts by merging the Provincial Education Assessment System in Punjab with PEC and hiring new staff. By November 2014, PEC had three such persons with PhDs.[7] PEC also hired five subject specialists for key subjects for grades 5 and 8 separately. However, the research and analysis department still lacks stable personnel. Thus, activities such as research or communication of findings and promotion of public discussion are still lacking.

PEC relies on temporarily contracted staff for item and rubric development as well as scoring. It draws its item writers from a pool of subject specialists, largely experienced teachers. Selection criteria of item writers have become more formal and consistent in recent years.[8] PEC selects personnel on the basis of experience with assessment, teaching, and familiarity with curricular SLOs. Candidates are first tested on assessment and content. This is then followed by an interview. Candidates are further whittled down after a performance review (i.e. after attending professional development workshops and developing items). The scorers are drawn from a pool of subject specialists. They are currently hired on the basis of academic qualifications but criteria for selection are under improvement. In the future, criteria may include teaching experience and performance on a recruitment test.[9]

In short, PEC has a stable set of technically competent staff which it has achieved by increasing the number of its permanent technical staff and investing in higher salaries. It still, however, lacks a career path for such staff which could serve as an additional incentive for them to stay at PEC.

SINDH

Political commitment, policy and legislation

Similar to Punjab, the standardized assessment efforts in Sindh emerged at a time when the Sindh Government had already developed a broad agenda for education reform.

The Sindh Education Reforms Program (SERP) I was initiated in 2007, and has continued under SERP II. The Education Minister provided the political will and Reform Support Unit (RSU) along with the Education Department provided the leadership and initiative for establishment of the Standardized Achievement Test (SAT). As noted by the World Bank documents, the SAT has wider and deeper ownership as compared to the sample-based assessments conducted by the Provincial Education Assessment Center (PEACE) in Sindh.

The Sindh Government chose a very different path from Punjab for implementing large-scale assessments. In lieu of establishing legislation, a PC-1 was developed, which enabled them to have the SAT up and running quicker than it would have taken to establish the legislation. This may very well have served its purpose for the SAT's initial phase, however, moving forward the SAT will need to be taken out of project mode and given legal status to ensure its permanence. Thus, the current deliberations over the future of the SAT are significant.

Similar to Punjab once again, although there is an educational reform agenda, Sindh lacks an assessment policy. However, the Sindh Government has taken some steps towards addressing this. In 2015, the Sindh Government notified the Sindh Education Student's Learning Assessment Framework, the purpose of which is to address major shortfalls in the assessment system. What emerges is yet to be seen.

With regard to PEACE in Sindh, there is no legislation and there appears to be limited political will to support sample-based assessments conducted by it in the past.

Support of development partners

The World Bank has been supporting the Sindh Government's reform agenda since its inception and most recently it has been providing budgetary support to SERP II through the Second Sindh Education Sector Project (SESP) II along with co-financing from the EU.[10] Under this project, the SAT is a Disbursement Linked Indicator (DLI). The World Bank provides review, technical assistance, and advisory support to the SAT through a third party evaluation of each round of key aspects of the SAT.[11] Currently, the SAT is being fully funded by the SESP II.

During the period in which PEACE in Sindh began its district representative assessments, it received technical and financial assistance from development partners. Its activities were a DLI under the World Bank supported Sindh Education Sector Project (SEP).[12] The EU provided technical assistance as well. However, this was discontinued under SERP II as discussed in Chapter 2, because of lessons learned by the World Bank in SEP and because RSU and the Education Department had prioritized the large-scale SAT over sample-based assessments conducted by PEACE moving forward.

Institutional arrangements

Rather than establishing a new institution, as in Punjab, or working with an existing one within the public sector, as in Khyber Pakhtunkhwa (KP), Sindh chose to outsource its large-scale assessment to a private sector entity. The Sindh Government indicated that, "contracting testing services from the open market through a competitive, transparent, and objective process potentially improves the quality of the activity."[13]

As discussed in Chapter 2, Sukkur Institute of Business Administration (IBA) won the bid to implement the SAT in Sindh. Design, conduct, marking, and dissemination of SAT results all fell within the mandate of Sukkur IBA. The main objectives of the SAT are: linking educational reform to outputs; using results to change teacher training, recruitment policies, and the pedagogical practices of teachers; reviewing the curriculum and syllabus; and shifting to a results-based accountability system where parents and other stakeholders are informed.[14] Sukkur IBA was also responsible for developing an item bank.

The entire SAT activity from 2012 - 2016 was funded through a DLI of the World Bank SESP II. RSU was responsible for managing the funds allocated by the World Bank for the SAT activity. The World Bank is providing funding for another round of testing- the fifth round of the SAT- as part of a one-year project from 2016 - 2017. RSU will invite bids soon and Sukkur IBA will probably respond to this call since it has developed the expertise to implement this test over the last few years. Once the World Bank funding exhausts, the expectation is that the SAT will be funded by the Sindh Government but whether it can continuously fund the SAT remains a question.

PEACE Sindh is located in the Bureau of Curriculum. Regular funding for PEACE comes from the provincial government. However, this funding does not support research and development activities in PEACE. Lack of funding has had a limiting effect on its activities.

Institutional structure and human resource

Sukkur IBA has a SAT project office with a core team comprising of a task leader, project coordinator, project manager, quality assurance manager, two monitoring officers, a program officer, 11 project officers, a web manager, and graphic designer. Apart from this, Sukkur IBA hires more task-based

personnel such as item writers (subject experts), item reviewers, web developers, translators, assessors, evaluators, and field-based administrative staff.

Item writers are selected on the basis of holding a Master's degree and having teaching experience. Initially, item writers came from the open market. Currently, the majority of item writers come from Sukkur IBA's five community colleges where teachers are hired from Aga Khan University- Examination Board (AKU-EB) affiliated schools. Once selected, item writers are provided training by experts on various aspects of assessment design, for example how to select appropriate language passages and how to formulate credible multiple choice questions. Item writers are paid per item. Scorers are also hired from the open market. Sukkur IBA initially selects approximately 200 persons and hires about half for the task. These scorers receive a three-day training in which they practice marking and have a follow-up session to review problematic areas from the exercise.

Initially Sukkur IBA lacked technical capacity, particularly, in item writing. It worked with AKU-EB in the second year of the project to develop this capacity. Collaboration between community college teachers and the in-house team has also developed Sukkur IBA's capacity over the past three years. While staff is more familiar with item writing procedures, their capacity to conduct item analysis needs to be improved further.

PEACE in Sindh is housed in the Bureau of Curriculum and Extension Wing in Jamshoro. PEACE staff consists of the coordinator (who reports to the Director of the Bureau), subject specialists, research officers, and account officers.[15] PEACE suffers from lack of technical staff, psychometricians, and statisticians. As a result, its previous sample-based assessments were perceived as lacking for a number of reasons including that "no final review of assessment instruments takes place".[16] Analysis and reports for the 2009 and 2010 sample-based assessments had been completed by PEACE and disseminated to stakeholders such as the Education Department, RSU, Provincial Institute for Teacher Education, and the Sindh Textbook Board. However, analysis and reporting for subsequent assessments did not take place due to the limited capacity of PEACE in Sindh.[17]

PEACE is a gradually receding institution. The Sindh Government no longer appears to have the appetite for sample-based assessments. Furthermore, development partners are also supporting large-scale assessment in Sindh. Given its reduced role, this report will not discuss PEACE in the remaining chapters.

KHYBER PAKHTUNKHWA

Political commitment, policy and legislation

Assessment reforms are central to the overall reform agenda of the provincial government to ensure implementation of a uniform curriculum across the province.[18] The need for a centralized assessment system providing data on student achievement before grade 9 was first mentioned in the KP Education Sector Plan 2010 - 2015. Since then, the KP government has embarked on developing this assessment system. Currently there appears to be significant political will for developing this assessment system. However, since it is still very early in the process there is no indication of when legislation backing this assessment system will pass. Finally, similar to other provinces, there is no clear assessment policy while clearly one is needed, to resolve the issue of which kinds of assessments should be conducted (keeping in mind that the sample-based assessment still exists in KP) and at what level.

Support of development partners

DFID through its KP Education Sector Program has supported the KP Government's education reform agenda since 2012.[19] Under this program, it has supported thinking and development of different options for developing a standardized assessment system. It has commissioned studies to explore the options available within the province. DFID is providing direct technical assistance to organizations involved in assessment such as Provincial Education Assessment Center (PEAC) (particularly for analysis and dissemination of results) and the Boards of Intermediate and Secondary Education (BISE). Apart from DFID, Australian Agency for International Development and the EU are also providing budgetary support to the education sector in KP.[20]

Institutional arrangements

The KP Government has taken yet another path towards institutionalizing large-scale assessment in the province. The BISE in KP will be playing a larger role in organizing large-scale assessments in the province. This will include designing, administering, scoring, and communicating the results of assessments to all stakeholders. Many challenges exist in doing so, not least, the BISE's own reluctance to take on this responsibility. Moving forward, the government will need to address this along with lack of staff, capacity, and budget to carry out activities. The BISE will need a new mandate to undertake primary and elementary level assessments and appropriate funding to fulfill such a mandate.

The sample-based assessment in KP is managed by PEAC, which is housed in the Directorate of Curriculum and Teach-

er Education. It was established under a PC-1 and continues to run in project mode. PEAC's mandate is to plan, design, and conduct sample-based assessments as well as to build assessment culture in the province. However, there remains lack of clarity regarding PEAC's future direction (as discussed in Chapter 2). It also lacks stable funding which hinders its activities. For example, a shortage of funds meant it could not disseminate the results of its 2015 assessment efficiently.

Institutional structure and human resource

Currently, the BISE do not have adequate capacity to organize the standardized large-scale assessment. A roadmap for improvement of the BISE has also been developed which focuses on three main areas of change: (1) human resources- new technical staff will be added as required at all levels, including item writers, data analysts, and psychometricians; (2) further capacity development of existing staff will take place; and (3) infrastructure development-investments in equipment as well as software for scoring and monitoring will be made. Item writing workshops, under the KP Education Sector Program, have already begun with BISE staff for developing items for the large-scale assessment as well as the grades 9 - 12 examinations.

PEAC is currently providing assistance to the BISE in the design of the grade 5 assessment. The current PEAC team comprises a deputy director, a systems analyst, three subject specialists, and additional administrative staff. The current PEAC team is no longer technically strong. It has lost some key staff that had received substantial training from the British Council and National Education Assessment System (NEAS) and developed experience through implementing various rounds of testing. In addition, there are, at present, no personnel for the analysis of data either. In 2015, for the sample-based assessment, analysis had been outsourced to a consultant with experience working at the World Bank. A restructuring of PEAC is planned. The Secretary of Education has taken an interest in restoring former PEAC team members to the institution as this will be beneficial to strengthening the assessment culture in KP. Whether this will happen remains to be seen.[21] Government rules determine transfer and posting of staff at PEAC; when officers are promoted to Grade 19 they are transferred. It is important that even if these ex-team members are not restored, persons with appropriate qualifications and expertise are placed within PEAC to contribute to its growth.

SECONDARY AND HIGHER SECONDARY LEVEL EXAMINATIONS

BOARDS OF INTERMEDIATE AND SECONDARY EDUCATION

Political commitment, policy and legislation

The BISE were established at a time when national education policies reflect a need to reform secondary education and were envisioned to manage all aspects of this critical stage, of which examinations were one aspect. The documents recognized the role of examinations in transforming teaching and placed an emphasis on both external examinations as well as internal school-based exams. There was, therefore, commitment for examination reform, albeit different from what happened in practice. The Government chose to establish the BISE as statutory bodies, formed on the basis of an act authorized by the provincial legislatures.

The BISE became solely associated with their function of organizing external examinations. From very early on, the policies recognized the need to improve the examination systems and institutions. A task force appointed by the Ministry of Education in 1985, noted the need for alignment between test content, curriculum objectives, and teaching/learning processes. The education policies of the 1990s sought to improve the capacity of examination staff, mechanize the process of preparing and declaring results, redo the format of examination papers to include objective type, short answers, and essay type questions, and discourage rote learning. The National Education Policy 2009 and the consequent education sector plans echo many of these points in relation to assessments. The policy specifically discusses reducing differ-

ences in quality between the examinations conducted by the BISE by reducing the number of such boards. Despite these policies there has been very little political will for reforming the BISE. At most, the emphasis has been on attempting to curtail cheating, while improving practices related to design and use of examination data remains outside the realm of efforts made.

Support of development partners

The development partners have provided virtually no support to the BISE. The lack of such support is in large part due to the lack of attention paid to secondary education by development partners who have traditionally focused on the primary and elementary stages of education.

Institutional arrangements

As mentioned previously, there are currently 28 BISE in Pakistan. Their proliferation has been justified on the basis of increased numbers of schools and student population. However, given their mandate, this does not appear to be warranted. In addition, as the case of India indicates, one board per state or province should be sufficient.

The BISE are autonomous organizations. They have a working relationship with the respective departments of education and are overseen by the chief executives of the province. There is a document known as the Calendar for each board that governs the functioning of the boards in minute detail.

Inter-and intra-provincial coordination committees coordinate the functioning of the BISE. The Inter Board Committee of Chairmen (IBCC) consists of chairmen from all the BISE. Its responsibilities include setting standards and ensuring a measure of uniformity in academic and evaluation standards amongst the BISE, providing a forum for discussing and debating issues related to intermediate and secondary education, granting equivalence to foreign certificates, and attesting certificates and diplomas. In practice, the IBCC does not play a significant role in setting standards or reviewing practices. The intra-provincial committees such as the Punjab Board of Committee Chairmen make province specific decisions such as those related to the choice of paper setters or selection of examiners. After the 18th Amendment, these provincial bodies are supposed to expand their role. However, they have not so far.

The BISE Act in each province ensures a certain degree of similarity in the mandates of the BISE across the provinces. The acts provide clarity and guidance on the composition and powers of the boards, delegation, budgeting, and auditing in addition to general rules and regulations. As mentioned in Chapter 2, the BISE have a mandate beyond examinations. Yet, in practice, their efforts are concentrated on organizing examinations at the secondary and intermediate levels. The BISE focus on certification and promotion. It is significant that the mandate of the BISE does not include use of examination data for the purposes of research and feedback to the education system at large. Funding for the BISE comes from the fees they charge students.[22]

Institutional structure and human resource

All the BISE are structured in the same way. There are two main divisions: one is administration and finance and the other is examinations along with an audit department and Office of Confidential Press (refer to Figure A in the Appendix for the BISE organogram). The examinations department organizes all activities in an examination cycle as well as other exam related activities (affiliation and registration of schools, enrollment and registration of students, development of syllabi and model papers). In some boards (Peshawar, Lahore, and Karachi), there is a computer cell that keeps computerized records and tabulation of marks. In some cases, there is also a research and development cell. However, this cell mostly undertakes non-research activities. A clearly identified department that deals with exam paper construction does not exist in any of the BISE.

Almost all of the BISE lack technical and professional staff with skills to effectively design and score examinations, and analyze data generated. All paper development and scoring activities are outsourced and research staff is virtually non-existent.

Paper setters, head examiners, and examiners are public school teachers with considerable experience either from secondary schools or colleges with appropriate subject specialization. Paper setters are usually selected from an existing pool of persons. The recruitment criteria is similar across the BISE, with some exceptions, such as Karachi Board of Secondary Education (BSE), where examiners must have five years teaching experience at the secondary level along with a Bachelor of Education or Master's level degrees. There is also a position of a moderator, a very senior subject specialist, who is responsible for reviewing and finalizing the papers at the Karachi BSE.

The lack of capacity of paper setters and examiners is a significant issue. The flawed recruitment process is considered to be one of the main reasons behind this capacity deficit; under-qualified persons get selected due to lack of adequate criteria and transparency. To address this issue, some efforts have been made to improve selection processes. For example, in Punjab an initiative led by the Chief Minister in 2010 sought to develop a pool of best paper setters. Accordingly, the BISE devised a set of criteria that gave appropriate

weight to academic and professional qualifications, professional experience, and a recruitment test. On the basis of this process, ten persons for each subject were selected for setting papers. Similarly, a recent effort has been made in KP to create a database of professionals to enable a better review of credentials of potential paper setters.

Another reason for poor capacity of paper setters and examiners is lack of adequate training. There is no program for capacity building, and when it does occur, it seems most professional development activities are conducted on an ad hoc basis. For example, in 2013, the Karachi BSE conducted a ten-day training of paper setters and the Peshawar BISE also trained 48 paper setters. However, these are not regular efforts.

With regard to analysis and research systems, analysts and programmers can be found on the staff of some boards. However, they do not work on data analysis. Rather, they simply process data to produce the exam results. There are positions for research in some boards, for example, in Lahore and Karachi. However, for the most part, these positions are either unfilled or filled by individuals with no research skills.

AGA KHAN UNIVERSITY-EXAMINATION BOARD

Political commitment, policy and legislation

In the case of AKU-EB there has certainly been leadership at the level of Aga Khan University to establish the board and design and conduct examinations of a superior quality. The examination also has legislative cover in the form of an ordinance, thus ensuring its ability to operate and cater to students. However, AKU-EB still remains at the margins, unable to cater to mainstream public education. There is virtually no political interest in promoting the AKU-EB examinations. Not only have provincial governments not improved the existing BISE, it appears that they have also not taken any steps to ensure schools can affiliate with AKU-EB.

Support of development partners

Development partners only played a role in financing AKU-EB (for details refer to the next section). Apart from this, AKU-EB made use of its own network to provide capacity development to its staff.[23]

Institutional arrangements

As stated in Chapter 2, the 2002 Ordinance recognized AKU-EB as a fully autonomous and self-regulatory institution with the freedom to offer examinations that follow the national curriculum up till the secondary level to students in the private sector. Originally, the Ordinance stated that government schools under the Federal Government could opt to affiliate with AKU-EB and left terms in the provinces up to the provincial governments. However, as stated in Chapter 2, the board later agreed not to seek affiliation with any government schools. The Ordinance recognized AKU-EB as a member of the IBCC with the freedom to engage in collaborations with the IBCC.

The Ordinance called for the establishment of a board of directors of AKU-EB, to be supervised by Aga Khan University's Board of Trustees. The Chairman of the IBCC, or the Chairman's nominee, is a member of the board of directors and the Director of AKU-EB, appointed by the University's Chief Executive Officer, is a member of the IBCC.

Upon establishment of AKU-EB, the Aga Khan University's Board of Trustees identified an aggregate budget of $7.3 million for the developmental phase from 2003 - 2007. Funding possibilities with various donor agencies were explored.[24] In 2002, United States Agency for International Development, as part of its support to the Education Sector Reforms Action Plan, contributed $4.5 million toward AKU-EB's budget.[25] The remaining $2.8 million was met by Aga Khan University. Only $5.66 million of the total grant was actually utilized because of the one-year delay in launching full-scale operations and enrollment figures being lower than originally predicted.[26] The amount not utilized was allocated to meet the operating shortfall of subsequent years.

The sustainability of AKU-EB's financial model and the associated figures were initially derived from the assumption that public sector schools would play an essential role in providing candidates and generating revenue.[27] After public sector schools withdrew their affiliation because of the reasons underlined in Chapter 2, the revenue generated by AKU-EB was not enough to cover the operating shortfall till 2008.[28] Exact data for subsequent years is not available. AKU-EB's commitment, backed by Aga Khan University's broader vision, not to substantially raise fees for students (to ensure that examinations continue to remain accessible to as wide a demographic of students as possible) has delayed AKU-EB's move towards self-sustainability.

When AKU-EB started offering examinations in 2007, a tiered fee system was established to ensure broader access.[29] For the Secondary School Certificate, students from schools with a monthly fee of PKR 800 or below paid PKR 1,530 for the two-year examination. Students from schools with fee levels above PKR 800 were charged PKR 3,060 (with students who took examinations with a practical component paying a little extra).

Institutional structure and human resources

AKU-EB consists of five main wings, that are overseen by the director: Assessments, Curriculum and Exam Development, Teacher Development, Operations and Business Development, and Communications (refer to Figure B in the Appendix for the AKU-EB organogram). The curriculum and exam development unit focuses on the revision of syllabi and exam design, encompassing the whole process from item writing to the development of the final examination papers. The exam design processes are supervised by a core team of subject specialists, who are experienced teachers in their relevant disciplines with additional training and experience in assessment design.[30]

As conduct and scoring are seasonal activities, AKU-EB hires part-time staff on a cyclical basis. These positions, along with relevant experience and qualifications being sought, are listed on the AKU-EB website. Applicants can register their interest through email or by filling out application forms that are available online. These positions include subject panelists– responsible for developing and revising syllabi and recommending appropriate textbooks while considering the objectives of the national curriculum; item writers– who develop appropriate items after attending a three-day workshop; translators; e-markers– responsible for scoring papers under the supervision of senior staff; visiting examiners– responsible for observing and scoring candidates' performance in practical examinations by visiting AKU-EB affiliated schools across the country; and exam supervisors and invigilators– who act in a supervisory capacity at examination centers when exams are in session.

Teachers interested in becoming item writers are invited to apply and shortlisted candidates are called for interviews. Applicants are required to have teaching experience in the relevant subject and grade level and must possess a post-graduate degree or equivalent to be considered eligible.

CONCLUSION

The foregoing discussion shows that the degree of political commitment varies for different types of assessments. There is a clear commitment amongst the provinces for establishing large-scale primary and elementary level assessments and to use the results of such assessments to improve education service delivery. The focus of provincial governments on large-scale assessment is driven, amongst other things, by a political desire to implement a common core curriculum to all students regardless of the modality of service delivery. As a result, the need to develop a standardized measure of student achievement has emerged. Development partners' interest is driven by a need to support the government in its own efforts to improve quality education as well as to generate evidence of the effectiveness of their support to the education sector.

While there seems to be a lot of focus on primary and elementary level assessments, secondary level assessments, although well-established, do not receive a similar level of government and donor commitment for reform. The reforms, where they have taken place, are too small and only in the private sector. AKU-EB, which was established to address many of the limitations of the BISE examinations, continues to remain unable to cater to the public sector and no efforts have been made to change this by the governments.

Legislation and policy need to go hand and hand. In some cases one is racing ahead of the other. For example, Punjab has been quick in providing supporting legislation to large-scale assessments. However, the results of the examinations conducted by PEC have not been used to drive improvements in teacher training. Apart from PEC, other institutions, such as the Directorate of Staff Development and Punjab Education Foundation, have been conducting their own assessments in the province. This situation has resulted in assessment overload in schools. The need for a coordinated assessment policy continues to exist. Policy has lagged behind legislation in Punjab. In KP, there is an emergent policy to use large-scale assessments to drive improvements in the education system. However, KP has no legislation on the large-scale assessment yet. Hence, no institution has the legal mandate to administer large-scale assessment yet. Legislation is lagging behind policy in KP. Sindh has neither the legislation nor a well-defined assessment policy. The level of seriousness about assessments is indicated by the establishment of legislation for assessments agencies- the BISE, AKU-EB and PEC are all mandated by law. Only Sindh has yet to decide which path it will take with regard to the SAT.

In Sindh and KP, the provincial educational assessment agencies that were created in the past in tandem with NEAS are still active. At the national level, NEAS has re-emerged after being dormant for a long time. However, neither the federal government nor the provincial governments have an assessment policy detailing which assessments will be conducted at what level, how results will be used, and for what purposes.

Such a policy, once formulated, would help streamline efforts, ensuring greater efficiency in utilization of limited human and financial resources available for assessment activities.

Legislation and policy are necessary but not sufficient conditions for high quality assessments. Human resources matter. As can be seen in the cases of Brazil and Uganda, without investing in human resources, assessment practices will not be up to standard. While Pakistan has embraced the promise of modern systems of assessment, it is still catching up when it comes to human resources. This lack of human resource is most pronounced in the BISE which, given their mandate, are not structured or resourced as assessment agencies. The dearth of human resources can be linked to lack of noteworthy professional programs being offered in assessment at institutions of higher education.

Given the highly technical nature of work required in an assessment agency, it is important that it is led by individuals who have expertise in the field of assessment. This is not the case in some assessment agencies where leadership remains a bureaucratic post and not a technical position. This is likely to have implications for the leadership's ability to understand the need for certain technical processes in assessment design and limit the likelihood of their setting priorities for hiring the requisite technical staff needed to meet accepted assessment standards.

Lastly, assessment data use is at best limited in Pakistan. This is partly due to poor perception of the quality of some assessments, lack of capacity in assessment agencies to produce meaningful reports in a timely manner, lack of a communication strategy, and disconnect from the policy process. It is no surprise then that assessment results have had little impact on policy and practice (this issue will be explored in detail in Chapter 6). As found in the case of Brazil, regular and timely production of assessment results encouraged the development of an assessment culture in which there was greater demand and use of assessment results.

This review of the enabling context for large-scale assessments in Pakistan highlights the need for coordinated developments in levels of legislation, policy, and human resource. The children of Pakistan deserve high quality assessment services. Politicians, policymakers, and academic institutions need to, therefore, act in a coordinated manner to identify and address gaps in the enabling context.

[1] Clarke (2012)
[2] Ferrer (2006)
[3] Ibid.
[4] Ibid.
[5] DFID (2014), p.9
[6] I-SAPS (2015)
[7] DFID (2014)
[8] PEC staff interview (6 November 2015)
[9] Ibid.
[10] As part of SESP the EU has provided 25.5 million Euros for the period 2012-16.
[11] World Bank (2013). In practice, according to interview respondents, the third party evaluation was conducted by PEACE and education department staff as well as World Bank officials who are also involved in the SAT through the working group. The review is, therefore, in practice an internal review. According to respondents, the findings of the review are positive and not very critical.
[12] According to World Bank (2012a), In 2009 (math test for grade-4), 2010 (language test for grade-4), 2011 (follow-up math test for grade-4, science test for grade-4 and math test for grade-8) and 2012 (follow-up language test for grade-4, social studies test for grade-4 and language test for grade-8)
[13] World Bank (2013), p.42
[14] RSU & Sukkur IBA (2012)
[15] "Bureau of Curriculum Structure" (n.d.)
[16] World Bank (2012d), p.7
[17] World Bank (2012a)
[18] BISE Peshawar staff interview (26 February 2016); PEAC staff interview (26 February 2016)
[19] KP Education Sector Programme has contributed 203.5 million pounds for the period June 2012 to October 2016 and an extension till July 2020 and an additional 79.7 million pounds had been requested and granted as well.
[20] Australian Agency for International Development has contributed 41.3 million pounds expected to last till 2016 and the EU has contributed 35 million pounds for the period 2015 – 2018.
[21] PEAC ex-staff interview (26 February 2016)
[22] The amounts of fees vary depending on which exam (matric or intermediate), science or arts subjects and whether they are private candidates and the board itself.
[23] USAID (2008)
[24] Saleem (2007)
[25] USAID (2008)
[26] Ibid.
[27] AKU-EB former staff interview (11 March 2016)
[28] USAID (2008)
[29] Ibid.
[30] Ibid.

04

ASSESSMENT DESIGN PRACTICES

INTRODUCTION

Key characteristics of good assessment design include validity, reliability, and equity/fairness. An assessment is considered valid when it tests what it has intended to test. It is considered reliable when the scores or results are comparable over time for different test taking populations. It is equitable when it meets requirements of fairness, preventing bias of any form in the assessment design. Designing assessments with these characteristics is a highly technical process, requiring several months to a year of work from the development of an assessment framework to test finalization.

This chapter describes standards and practices for assessment design practices. It reviews the extent to which existing practices in Pakistan are aligned with the accepted standards for designing high quality assessments. For every case, each stage of the assessment design process will be reviewed including test specifications, training and item development, item pilot and psychometric analysis, and technical documentation.

STANDARDS AND BEST PRACTICE[1]

The practice of design of good assessment instruments has evolved into a highly refined craft backed by advances in psychometrics (i.e. the science of psychological measurements). A typical design cycle involves contribution from curriculum experts, subject specialists, and psychometricians. Assessment design is governed by a document, typically called the test specifications, that specifies the contents of the test. Item development can only begin after the test specifications have been created. The items are written, reviewed, and pilot tested for their validity, reliability, and psychometric robustness by teams of subject specialist reviewers and psychometricians. This involves determining the alignment of different items with the curriculum and their difficulty levels. The process must result in an assessment instrument that can reliably distinguish between the abilities of test takers and validly represents the curriculum content that it intended to test (refer to Figure 4.1 for an overview of the assessment design process).

Assessment framework and test specifications

An assessment framework is usually developed to identify the purposes of the assessment and assessment standards often linked to agreed-upon goals. This framework defines, specifies, and describes the content and skills to be assessed (often guided by a curricular document), the target population, the type of assessment (norm referenced versus criterion referenced), the manner of analysis, and the nature of reports based on results. The framework is developed in collaboration with key policymakers and stakeholders outside of the assessment agency to ensure broad based support for it and its integration with other quality aspects.

As part of the assessment framework, a test specifications document or blueprint is developed for each subject (refer to Table 4.1). It details the curriculum and content areas as well as cognitive skills (knowledge or recall, interpretation, and application) to cover, the question format such as Multiple Choice Questions (MCQs), Constructed Response Questions (CRQs) or Extended Response Questions (ERQs), the response format, the scoring procedures, the test length, the desired psychometric characteristics of items (such as difficulty and discrimination), as well as other test characteristics such as reliability.

Figure 4.1 Overview of assessment design process

Table 4.1 Distribution of math items by cognitive domain TIMSS grade 4
Source: Adapted from TIMSS (2003), p. 343

Cognitive domain	Percentage of items	Total number of items	Number of multiple choice items	Number of constructed response items	Number of score points
Knowing facts and procedures	23	45	35	10	45
Using concepts	19	37	31	6	39
Solving routine problems	36	70	43	27	76
Reasoning	22	42	19	23	55
Total	100	194	128	66	215

The difficulty of the test depends on its purpose. If it is meant to be administered to all students in a given population, then two-thirds of the test should consist of items that are medium difficulty (this means that two-thirds of the population have between a 30 and 70 percent likelihood of answering the items correctly) and the remaining should be evenly divided between items that are easy (more than 70 percent of students are likely to answer these correctly) and difficult (fewer than 30 percent of students are likely to answer these correctly).[2] With regard to validity, those developing the assessment instruments should clearly indicate a) how the scores should be interpreted and used; b) for which population; and c) the constructs (knowledge and skills) being assessed.[3]

Training and item development

What were called test questions in the traditional examinations are known as items in the lexicon of modern standardized tests. Items may be thought of as units of interaction in tests. Before the test is assembled, items must be written and tested for their robustness. Item development is an elaborate process and is described below in some detail. Typically, it begins with a training or orientation that is provided to item writers. The purpose of this training is to help item writers develop a) an understanding of the overall purposes of the assessment; b) a familiarization with test specifications; and c) knowledge of item writing principles. Given its scope, this training often takes the form of a hands-on workshop to write the items.

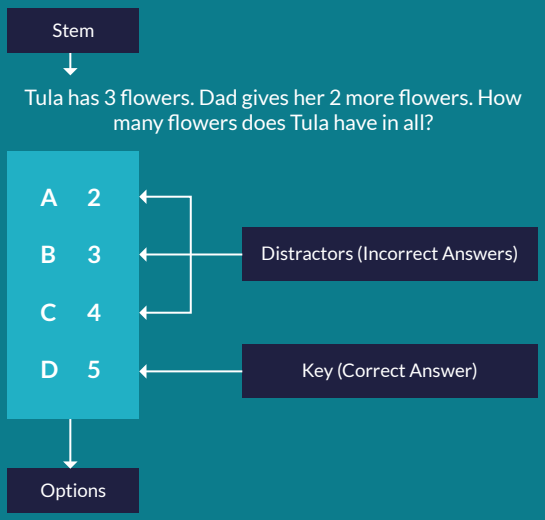
The item writing team, consisting of trained subject specialists, creates a large pool of items according to the test specifications. During the item writing process, item writers look out for the following characteristics: item difficulty, item bias or inaccuracies, item format specifications, and even item layout (refer to Box 4.1). The team also develops answer keys and scoring rubrics.

In the last stage, the item writing team, along with the subject and language experts, conducts a thorough review of the items. The items are reviewed with reference to the characteristics listed in Box 4.1 as well as for content, language, and phrasing. Items are then revisited and their quality improved. Often several rounds of item review are conducted before items are finalized.

BOX 4.1 CHARACTERISTICS OF GOOD ITEMS

- Can be mapped back to important item characteristics stated in the framework
- Fair (see glossary for detailed definition)
- Written in clear, simple language without convoluted or long sentences or unfamiliar terms
- Stand alone and do not depend on an understanding that has formed the basis of a previous item
- Preferably expressed in positive terms; negatives tend to cause confusion

The figure below shows the structure of a typical MCQ.



Good practices for MCQs

- **Avoid introducing grammatical or logical cues in the stem and key that point to the correct answer**
- **Avoid making the correct answer much longer or more detailed than the other options**
- **Minimize the amount of reading**
- **Avoid negative stems (if stem can only be expressed negatively, highlight 'not' by using bold type or italics)**
- **Vary the use of paired distractors so that patterns are not discernable**
- **Distractors should all be plausible but indisputably incorrect, while the key should be indisputably correct**
- **Stimulus material should be factually correct, self-contained, should be of interest to the target audience, and should not contain superfluous material**

Good practices for CRQs and ERQs

- **Clearly communicate the response students are expected to provide**
- **Ensure that short-response questions have more than two possible responses (e.g. if 'open' and 'shut' are the only options, the student has a 50% chance of being right)**
- **Quoting from the stimulus is preferable, rather than summarizing or interpreting meaning as this does not encourage close reading**
- **Instructions such as 'explain your answer' or 'give reasons for your answer' are preferable when there is a danger that students may give a superficial response to an open-ended question**

Source: Adapted from Anderson & Morgan (2008)

Item pilot and psychometric analysis

In traditional examinations, one person alone is usually responsible for setting the entire paper. However, in the context of modern day tests, significant effort is required to write, test, and finalize items that will appear in the test as items need to be representative of student ability, with their difficulty level being determined by the number of students who are able to answer them correctly. Item piloting is an essential step in identifying items that can be used when assembling equivalent, reliable, and valid tests. Item piloting entails testing a larger number of items than what is found on a test on a sample group with similar characteristics to those who will be taking the test. It also entails piloting the scoring schemes and rubrics. Empirical data is generated and a statistical and psychometric item analysis is conducted to determine which items are robust.

There are two main techniques or theories that are used to analyze assessment data: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT and IRT both allow for item-level analysis including the difficulty and discrimination of each item. However, the item statistics produced by CTT are not independent of test taker characteristics, while IRT produces item statistics that allows for the item to be characterized independently of the test taking sample.[4] Thus, IRT is commonly used in international assessments and perceived as being more applicable to large-scale assessments.[5] How-

ever, CTT is considered easier to use as its application does not require an advanced knowledge of statistics. Using IRT requires a certain level of skill which is not always available within country.

Aspects taken into account when selecting items are congruence with test specifications, difficulty level, discrimination index, whether the inclusion of the item improves the test, and whether the item bias is within acceptable limits. High quality items that meet these standards are placed in an item bank from where they can be retrieved for reuse with minor changes. Once the items have been selected, the tests are finalized, formatted, proofread, and printed.

Items corresponding to similar curricular goals may be different but can at the same time be equivalent. That is to say, the different items validly measure the same set of knowledge and skills. If the difficulty level and curricular objectives of a particular item or a particular collection of items is the same, the tests can be said to be equivalent to each other. They may, in fact, be seen as different versions of the same test. Such equivalence allows for comparisons to be made between individuals and groups taking different versions of the same test.

Technical documentation

Documentation of an assessment includes a description of the purpose and uses of the assessment, test specifications, item formats, scoring procedures, technical data such as psychometric indices of the items, evidence of validity and reliability, and cut scores or rules for interpreting the data.[6] The purpose of such documentation is for users of the test to have sufficient information to make sound judgments about the nature and quality of the test and to understand interpretations based on test scores.

PRIMARY AND ELEMENTARY LEVEL ASSESSMENTS

The following section will provide case studies of the design process for primary and elementary level assessments in Pakistan.

PUNJAB

The Punjab Examination Commission (PEC) exams are high-stakes and census-based. They are administered annually to all students in public schools and selected private schools across the province of Punjab in grades 5 and 8 in key subjects.

Framework and specifications

Since the 2012 exam cycle, PEC has developed test specifications for each subject to guide item development. The frameworks assign weight to items on the basis of content, Student Learning Outcomes (SLOs), and learning levels or skills. PEC continues to update and improve its assessment framework with donor assistance.

The PEC exam covers five main subjects including Urdu, English, math, science, and Islamiat. The papers are developed in Urdu and English. The exams consist of both MCQs and CRQs. The format of the paper has recently changed with the proportion of CRQs being increased from 40% to 50% as a means for prioritizing higher order thinking skills and writing skills that are not as easily tested in MCQs. PEC is keen to make an even greater proportion of the assessment CRQ-based.

The exam covers approximately 70% of the curricula, which is fairly high. The skills tested in the PEC exam at present are knowledge (30% of items test this), understanding (50% of items test this), and application (20% of items test this). Previously the PEC exam did not test understanding and application; it is only recently that it has begun to do so. Proportions have been assigned according to associated competencies in the curriculum. PEC has also attempted to shift from textbook-based items to curriculum-based items. For example, while some of the reading passages included in the PEC exam are lifted from the textbook, the questions associated with these passages are different to those included in the textbook. Ideally, PEC would have liked to use unseen reading passages in the exam (i.e. those not included in the textbook at all). However, PEC feels that doing so may overwhelm students and teachers.[7] The shift will have to be gradual.

BOX 4.2 PEC TEST PAPER REVIEW

A preliminary review of the current PEC exam papers (official grades 5 and 8 model papers, English and science 2016) suggests that their design has been informed by standard practices. For example, vocabulary questions were only asked in items that had an associated reading passage, so the meaning was contextual and not isolated; there were no grammatical or logical cues in the stem that pointed to the right answer and no differentiated length or structure between the correct options and the distractors; all negative stems had the word 'not' capitalized; and open-ended questions clearly communicated expectations for responses by providing instructions and prompts when necessary. In addition, questions did not appear to be drawn from the textbooks as is the practice in the traditional examinations administered by the BISE.

Training and item development

The item development process begins with subject specific item writing workshops. Shortlisted item writers attend a four to five-day workshop where they are briefed on the specifics of item writing such as alignment of items with SLOs, principles of assessment, and criteria for item review. After item writers are finally selected, they attend a refresher training workshop that addresses issues that arose in the first cycle of item writing. These workshops, conducted by both in-house and external resource persons, describe the various purposes of assessment, principles and practices of item construction, as well as construction of MCQs and to some extent CRQs.[8] Although the current training is perceived as an improvement from previous years, there is room for further improvement.

Item writing is guided by adherence to test specifications (particularly SLOs and difficulty levels) and characteristics of good items. The quality of items has improved over the years according to PEC and technical assistance staff (refer to Box 4.2). In addition, the wording of questions has improved to reflect thinking skills being tested (e.g. the verb used at the beginning of the question indicates the thinking skill being tested). Detailed rubrics have also been developed for each item, grade, subject, and version of the exam to make the scoring more consistent and valid across different marking centers.

After items have been developed, PEC subjects them to a review process. PEC now has staff to conduct both technical and content reviews. The items are first reviewed internally. In this internal review, each item is examined by a PEC subject specialist, assessment expert, and psychometrician. This is followed by an external review conducted by subject teachers. PEC has also appointed a translator to review Urdu and English versions as the PEC exams are bilingual.

Item pilot and psychometric analysis

Since 2014, PEC has begun to formally pilot its items and scoring rubrics. However, it is not clear how rigorous this pilot is as no details have been shared nor are the pilot results shared publicly in the form of a report. A statistical analysis of the items is conducted and the difficulty levels of items are established. PEC was using CTT to do conduct its analysis, however, in the last year PEC has begun to use IRT, which is considered more appropriate for a large-scale assessment.

PEC assigns a score to each item based on the SLO and difficulty level attached to it. Using this process, it assembles four different versions of the papers. PEC feels fairly confident that test papers are equivalent across the different versions. In addition, it appears that PEC has managed to ensure equivalence from year to year in the number of topics covered as well.[9]

Technical documentation

To date there has been no documentation of the technical processes that PEC utilizes to develop the exam. However, for the 2016 cycle, such a report is being developed and will be available to the public as well.

SINDH

The Standardized Achievement Test (SAT) is low-stakes and census-based. It is administered to all students in public schools and selected private schools across the province of Sindh in grades 5 and 8 in key subjects each year.

Framework and specifications

The assessment framework for the first three rounds of the SAT (i.e. the first three years in which the SAT was given), was developed by Reform Support Unit (RSU), with the assistance of the World Bank. This framework allowed for items to be developed from both the curriculum and the textbooks in use. For the fourth round of the SAT, Sukkur Institute of Business Administration (IBA) has developed an assessment framework drawing from the Trends in International Mathematics and Science Study (TIMSS) framework.

The SAT covers language, math, and science. The SAT is developed in the three main languages in use in Sindhi schools: Sindhi, Urdu, and English. For each subject and grade, the test specifications divide items across different cognitive do-

mains and content strands, and also indicate what percentage of MCQs and CRQs will be included in the test. Items on the test are aligned with the cognitive specifications laid down in the curriculum to the extent possible. In addition to the tests, Sukkur IBA develops background questionnaires for head teachers, teachers, and parents to understand factors that may contribute towards student achievement.

Training and item development

Once item writers are selected, an orientation is held for them. Training is provided on various aspects of assessment design, for example on how to select appropriate language passages or how to formulate credible MCQs. The item writing team then develops five to six times more items than are needed in the actual test. Efforts are made to include items from both the curriculum and the textbooks.

The item review is done by experts drawn from multiple institutions. First, the items are reviewed in validation workshops where external reviewers from Society of Pakistan English Language Teachers, the Math Association, and the Provincial Education Assessment Center (PEACE) in Sindh review the items. Previously, items had also been shared for review by experts within the RSU. However, this has been discontinued as, at present, there are no experts working at the RSU. Finally, the items are distributed amongst experts internationally for a final review of the language of the test.

Item pilot and psychometric analysis

Items are piloted by Sukkur IBA before the actual test. The pilot test is supposed to contain three to four times the number of items on the actual test. Four versions of the test are developed for piloting. The sample for the pilot is determined through stratified sampling. The sample includes rural and urban, male and female, and all types of schools. In the third round of the SAT in 2015, the items were piloted among 1,866 students from 40 schools in 3 districts in the northern part and 3 districts in southern part of Sindh.[10] Assessment experts look at the validity, reliability, and difficulty levels of the items and accept or reject them based on their facility value (the difficulty or easiness of an item) and discrimination index (ability of an item to discriminate between low and high achievers). Finally, two versions of the test are used in the actual SAT. There is no discussion on processes for ensuring comparability between the two versions in the technical documentation.

Sukkur IBA did struggle with item piloting initially and lacked the capacity to conduct item analysis. For example, in the first round of the SAT, it piloted items without an adequate strategy and sample design and in the second round items were re-piloted as the response rate was too low. However, with time, item piloting has improved. According to Sukkur IBA, about 60 - 97% of items in the third round of the SAT were accepted from the pilot.[11]

For the fourth round of the SAT, Sukkur IBA is also conducting a qualitative analysis of items by asking students to interpret the meaning of the questions asked. Findings have been useful. For example, it was found that students do not respond well to items that contain two prompts. This and other such findings are being used to improve the quality of items.

Sukkur IBA had created an online item bank containing approximately 30,000 items. These items were, however, aligned with the old curriculum (i.e. the 2002 curriculum) since, at the time, the 2006 curriculum had not been implemented in Sindh. With the implementation of this latter curriculum in the province in 2014, Sukkur IBA is now required to add 15,000 new items to the bank that are aligned with the 2006 curriculum. In addition, it has to review the older items contained in the bank and retain 10,000 of those that are aligned with the current curriculum. Sukkur IBA is also developing profiles for the various items and this information will also be added to the bank. For items added to the bank during the third round of the SAT, item development workshops were held at Sukkur IBA with assessment experts from the University's Department of Education leading the training. Items were developed by subject specialists from IBA community colleges and alumni under the supervision of these assessment experts who also reviewed the items.[12]

Technical documentation

Each year, Sukkur IBA produces a comprehensive publicly available report on the SAT. This includes documentation on assessment design and details of the test specifications, item pilot, and the criteria for item selection.

KHYBER PAKHTUNKHWA

The Khyber Pakhtunkhwa (KP) Government has launched an annual centralized assessment for all students across the province in grades 5 and 8. For now, this assessment has been conducted at the grade 5 level only. The first high-stakes centralized assessment in both grades 5 and 8 will take place in 2017.

At present, there is division of labor as far as design and conduct of this large-scale assessment at the primary level in KP is concerned. The assessment instruments are designed and developed by the Directorate of Curriculum and Teacher Education and administered by the Boards of Intermediate and Secondary Education (BISE). KP is in the process of improving these assessments to measure a wide range of knowledge and skills. In subsequent years, it seeks to align these

assessments with the curriculum and not with textbooks. KP is also in the process of introducing standard practices in assessment design, including a pilot of the designed items.

Apart from large-scale assessments, sample-based assessments are also being conducted in KP. The Provincial Education Assessment Center (PEAC) conducted its first independent sample-based assessment in 2015 in grade 2 (on students who had just entered grade 3) in math, English, and Urdu. At present, the frequency and the grade level at which PEAC will administer the sample-based assessment in the province is unclear.

In 2016, PEAC in KP developed a new assessment framework and test specifications with technical assistance supported by the Department for International Development (DFID). At present, assessment design is in process. Items have mostly been written in-house by PEAC staff and Regional Institute for Teacher Education faculty members. Where item writers are hired externally, they are assessment specialists with experience in item writing.[13] Workshops are held to write the items. Items written in these workshops are internally reviewed with the help of the technical assistance agency. After the item pilot, these will be reviewed once more internally. There is greater confidence amongst PEAC staff that these practices will ensure greater validity and reliability of tests.

SECONDARY AND HIGHER SECONDARY LEVEL EXAMINATIONS

The following section will provide case studies of the design process for secondary and higher secondary level examinations in Pakistan. They consist of the Secondary School Certificate examinations conducted in grades 9 and 10 and the Higher Secondary School Certificate examinations conducted in grades 11 and 12. Both sets of examinations are high-stakes.

BOARDS OF INTERMEDIATE AND SECONDARY EDUCATION

Guiding documentation

The exam development process at the secondary and higher secondary level does not follow the standards outlined earlier. There is no assessment framework or test specification document that is used by the BISE. The Examination Rules and Procedures part of the BISE Calendar sets out the processes for all aspects related to the examinations. However, a review of the calendar indicates that it mostly details guidelines for administrative procedures rather than paper development or analysis. Paper pattern (proportions in which different types of questions will be included in the examination) is set by the Inter Board Committee of Chairmen. Paper pattern is the same for all boards across Pakistan: 20% objective questions, 50% short questions, and 30% descriptive questions are included in the examination. This distribution is used by the BISE to develop model papers, which are shared with the paper setters as guides for their own paper development. For the most part, exams administered by the BISE are based on textbooks instead of the curriculum.

Paper development, review and finalization

Once nominated, the paper setters are called in to develop a certain number of papers. This practice is contrary to the generally accepted best practice of developing and testing individual items. While there are some criteria for this process, these are not similar to the best practices outlined earlier- procedures tend to focus more on maintaining secrecy rather than guiding item development.

The BISE use the officially designated textbooks and model papers as well as their own experience and subject knowledge as guides for developing papers. Paper setters are expected to ensure a good distribution of exam questions from all chapters in the textbook. Once developed, the papers are put through a review process. The scope of this review varies depending on whether the papers are being developed collectively by several BISE or independently.

BOX 4.3 BISE TEST PAPER RESEARCH AND REVIEW

Studies conducted point to the narrow range of skills tested and repetition of questions in examinations administered by the BISE. One particular study (Shah and Afzaal, 2004) reviewed the papers of two boards, Karachi BSE and Federal BISE for four years in English and biology and analyzed the coverage of questions according to cognitive levels. It found a very high frequency of knowledge-based recall questions as compared with questions dealing with understanding. There was not a single question that required test takers to apply their knowledge and skills in problem solving. The study also found frequent repetition of the same questions and limited content coverage. The study also indicates that there is a limited spread of questions across a few popular chapters each year.

A preliminary review of current BISE exam papers (Lahore BISE grade 9 papers, English and biology 2015) indicates that the content for questions is still taken directly from the textbooks. For example, the grade 9 English paper asked students to write a letter or a story, the material for which was available in the lone grammar and composition textbook approved by the Punjab Textbook Board. This means students need to only memorize these passages and reproduce them on the test to attain a good score. In addition, there are numerous issues with the quality of the items, for example, vocabulary is tested without a passage to refer to and other grammatical errors persist.

Boards in Punjab and KP appear to follow a similar approach to paper setting in an effort to standardize the process. The responsibility for developing papers in different subject areas is distributed across all the BISE within each province. This mechanism ensures a modicum of uniformity, if not standardization, across all BISE within the same province. After developing several papers for each subject, these papers are then shuffled and divided across different versions to maintain secrecy. The process followed in Sindh is slightly different as each board develops its own papers independently. In each board, three paper setters are appointed per subject. Thus, three versions of the paper are developed for each subject. A moderator then reviews these papers to ensure similar levels of difficulty and course coverage.

The process of paper development takes approximately two months. However, the papers are finalized, at most, a few days or even the night before the examinations in an effort to ensure secrecy of the papers.

Given the process followed, it is unlikely that examinations at the secondary level meet the requirements of validity and reliability (refer to Box 4.3). This is further substantiated by research studies that indicate substantive problems with the process.[14] The papers set by different boards are not comparable. Even papers set within the same board are not comparable across time for the same subjects. Study findings reaffirm problems with the quality of the exam paper. Many contain errors in subject content and technical construction. In addition, they focus on testing a narrow range of low-level thinking skills and are dominated by content of approved textbooks.

AGA KHAN UNIVERSITY-EXAMINATION BOARD

Framework and specifications

The stated aim of the Aga Khan University-Examination Board (AKU-EB) is to foster and evaluate critical thinking, problem solving, and comprehension skills by developing examinations that are based on the national curriculum. The specifications of each of the examinations were initially determined in the syllabi formulated during the material development phase from 2003 - 2007. These syllabi, publicly available on the AKU-EB website, are periodically reviewed and revised by the curriculum and examination development unit within the board. The last comprehensive revision took place in 2012 and another review process is planned in 2016. The development of syllabi is a consultative process coordinated by subject specialists at the board with some input from outside subject experts.[15] These syllabi are used by item writers while writing test items, and the teacher training unit as supplementary resources. They are also shared with teachers and students at AKU-EB affiliated schools.

The syllabi provide detailed instructions on the examination of each subject as follows: (1) Student Learning Outcomes (SLOs) by cognitive level and skill; (2) distribution of marks by question type- MCQs, CRQs and ERQs- across all topics/competencies; and (3) types of questions to be included in each section, allocation of time, and distribution of marks across sections (refer to Tables 4.2 and 4.3). The syllabi also provide the weights given to cognitive domains derived from the national curriculum, which serve as a guide in the development of the exam papers.

Table 4.2 Number of Student Learning Outcomes by cognitive Level
Source: Adapted from AKU-EB (2012), p. 60

Topic no.	Topics	No. of	SLOs			Total
		Sub-topics	Knowledge	Understanding	Application	
1.	Introduction to Biology	8	2	11	1	14
2.	Solving a Biological Problem	1	-	1	1	2
3.	Biodiversity	7	5	15	2	22
4.	Cells and Tissues	4	1	21	1	23
5.	Cell Cycle	4	1	21	-	22
6.	Enzymes	4	2	13	1	16
7.	Bioenergetics	4	4	15	1	20
8.	Nutrition	6	10	16	2	28
9.	Transport	**8**	8	24	-	32
	Total	**46**	**33**	**137**	**9**	**179**
	Percentage		**18**	**77**	**5**	**100**

Table 4.3 Allocation of marks across question types
Source: Adapted from AKU-EB (2012), p. 61

Topic No.	Topics	No. of	Marks			Total
		Subtopics	Multiple Choice Questions	Constructed Response Questions	Extended Response Questions	
2.	Solving a Biological Problem	1	2	4	-	6
1.	Introduction to Biology	8	9	5	7	21
3.	Biodiversity	7				
4.	Cells and Tissues	4				
5.	Cell Cycle	4	4	3	-	7
6.	Enzymes	4	2	2	-	4
7.	Bioenergetics	4	8	11	8	27
8.	Nutrition	6				
9.	Transport	8				
	Total	**46**	**25**	**25**	**15**	**65**
	Practical					10
	Total					**75**

Training and item development

A three-day item writing workshop is held for selected item writers. During the workshop, item writers are trained to write items in simple, unambiguous language and are told to avoid including multiple prompts in the same stem. Some of the item writers are also involved in the development or revision of syllabi and are familiar with the elements of content knowledge and cognitive categories associated with SLOs in the syllabi.

During the process, item writers assign attributes to items that categorize their difficulty level and relevance along with the cognitive and content attributes already specified in the syllabi. These items, along with their assigned attributes, are reviewed externally by content experts. After the necessary modifications are made, items undergo a multi-disciplinary review in AKU-EB by two subject specialists as well as two to three specialists who are not from the associated discipline. For example, an item for the biology exam would be reviewed by two specialists from the natural sciences as well as two experts from the humanities. While the subject specialists examine the item for conceptual issues, the humanities experts assess the item's language for correctness and cultural appropriateness.

After the multi-disciplinary review, items and their attributes are fed into a question bank. They are retrieved when needed to assemble the exam papers based on test specifications. A map of all the item attributes is generated and items are selected according to the required attributes designated in the test specifications. Once the papers have been assembled, they go through several reviews. First, the papers undergo an editorial review. Second, an external expert, not affiliated with any of the AKU-EB schools, inspects the paper, re-validates the attributes, and determines whether it can be completed in the time made available to the students. Third, the manager reviews the exam paper followed by a review by the chief examiner. The director of AKU-EB then reviews the exam paper. Finally, an internal review committee inspects the exam papers to ensure there are no formatting inconsistencies or grammatical errors, spelling mistakes, or content issues. The committee is divided into two groups; each group checks a different set of exam papers and swaps them afterwards to ensure a thorough review. Scoring rubrics for questions are developed alongside the items and undergo the same review processes as the items. For a review of test papers refer to Box 4.4.

Item pilot and psychometric analysis

Once items are banked, AKU-EB pilots a representative sample of items on a representative sample of students from schools not affiliated with AKU-EB. With the permission of the school, the teacher development team teaches a topic and utilizes all the learning and assessment material developed for that topic. Afterwards, items developed for the same topic are piloted. Further details, however, on the statistical analyses conducted on the pilot data are not available due to lack of technical documentation.

AKU-EB typically prepares four versions of a particular paper. A fifth version is created if the paper is offered in Urdu. When the exam papers are finalized, they are coded and sent for printing. Exam papers are printed abroad to ensure secrecy and minimize the possibility of leaks. Each answer paper contains individualized student information on the front cover and each of the inner pages is bar coded.

BOX 4.4 AKU-EB TEST PAPER REVIEW

A preliminary review of papers (AKU-EB grade 9 and 10 papers, English and biology 2012) indicates that AKU-EB generally follows best practice in item development. In contrast to the BISE, AKU-EB uses source material from places other than the textbook. Reading passages are often taken from magazines, newspapers, and websites ensuring that the responses to the questions associated with these passages indicate comprehension ability. The example below illustrates this:

> **"However, it is believed that spices began to be used in ancient times, though not as specifically as they are used today. Their regular use actually started in the 10th century. Their multiple uses came to the knowledge of men when people in olden days used leaves from different herbs and shrubs for covering and keeping meat. The leaves would not only cover the food but also give a certain fragrance and added flavour to food. This interested them to research more about plants and seeds. And gradually they found different seeds, berries and barks, that later came to be used as spices. History tells us that Arabia was the hub of the spice trade. The Arabs were believed to be the ones who first introduced spices to European markets. The Portuguese and Spanish too were in the spice trade. The competition was tough in the spice trade and it is said that the Arabs enjoyed an edge in this trade."**
>
> **Source: Young World. DAWN**

CONCLUSION

The design of assessments bifurcates markedly, with more modern professional practices following accepted standards concentrated in the primary and elementary end of the assessment spectrum (refer to Table 4.4 for further details). The BISE, with the notable exception of AKU-EB, remain firmly ensconced in their decades long tradition of paper setting without recourse to best practices. This is not to say, that the rest of the assessment agencies do not need any further improvement. They too remain short of professional human resources needed to adhere to testing standards in letter and spirit.

PEC, Sukkur IBA, and AKU-EB have clearly developed test specifications, which they have been using for several years now. PEAC KP has also begun to use the standard best practices of assessment design in its sample-based assessments. These agencies train their staff for item development and follow, for the most part, suggested practices for item writing and review. The BISE, on the other hand, do not have any guiding specifications, nor do they train their staff or follow any of the required standards. They have a long way to go before transitioning and aligning with established standards. Item pilot and psychometric analysis appears to be one of the components of the assessment design process that requires the most work. Sukkur IBA has been conducting pilot tests since the inception of the SAT, however, it recognizes it has taken it a few years to improve the rigor and quality of the pilot test. Information on the pilot test sample, analysis, and results have been provided in its technical documentation. PEC has recently begun conducting formal pilot tests. However, the details of how they use pilot findings are not publicly available. Similarly, AKU-EB conducts pilot tests but provides limited details on the process and results. Technical documentation is also not available in the case of KP. For all cases, there is limited information available on the psychometric analysis.

While some of these assessment agencies have a good understanding of the standards for assessment design, the lack of publicly available technical documentation means that there is lack of evidence about which of the standards are actually being followed and how. Sukkur IBA is the only assessment agency that produces a publicly available report with technical documentation of the assessment. PEC will be developing one this year. In all other cases, the assessment design process is not publicly available. This is a major gap, as such documentation would allow for providing a clear picture of the design processes as well as the findings of the pilot and their implications.

Table 4.4 Assessment design practices in Pakistan

	PEC	SAT	KP G5	BISE	AKU-EB
Test specifications	✓	✓	×	×	✓
Training of item writers	✓	✓	×	×	✓
Item writing and review	✓	✓	✓	×	✓
Item pilot and psychometric analysis	✓	✓	×	×	✓
Item bank	×	✓	×	×	✓
Technical documentation (publicly available)	×	✓	×	×	×

[1] Anderson & Morgan (2008); American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999)
[2] Anderson & Morgan (2008), p.15
[3] AERA, APA, & NCME (1999), p.17
[4] Anderson & Morgan (2008)
[5] Carlson & Davier (2013)
[6] AERA, APA, & NCME (1999)
[7] PEC staff interview (6 November 2015)
[8] SAHE (2011)
[9] PEC consultant interview (11 March 2016)
[10] RSU & Sukkur IBA (2015)
[11] Ibid.
[12] Ibid.
[13] Several of whom are ex PEACE staff
[14] Bethell, Dar, & Crighton (1995), as cited in Christie & Afzaal (2005)
[15] USAID (2008)

05

ASSESSMENT IMPLEMENTATION PRACTICES

INTRODUCTION

For the purpose of this report, the term "implementation" refers to the administration and scoring of papers. It includes the development and implementation of Standard Operating Procedures (SOPs) for recruiting and training staff to administer and score the test, allocating test centers, distributing and collecting papers, administering tests, preventing use of unfair means, and scoring tests. Effective implementation requires a quality assurance or monitoring mechanism for ensuring adherence to SOPs and transparency in all practices. Most importantly, it entails "reasonable efforts" to eliminate opportunities for the use of unfair means in tests.[1]

This chapter will begin with describing best practices for assessment implementation. This is followed by a description and review of the implementation practices for assessments in Pakistan. In each case, the selection, allocation, and training of staff will be reviewed as well as the test administration and test scoring procedures.

BEST PRACTICE[2]

Selection, allocation and training of staff

Appropriate and competent staff is central to ensuring quality of test administration and scoring. Test administrators (or invigilators) are, more often than not, themselves teachers. However, in some countries test administration is done by university graduates or school inspectors and ministry officials. Where teachers are used, efforts are usually made to ensure that they are not teachers of students being tested, that they are from non-participating schools or are retired.[3] While not important for invigilators, adequate subject knowledge is of crucial importance for scorers.

Both test administration and scoring staff require appropriate orientation and training. Often administrators are provided with a detailed manual on administrative procedures and instructions. These manuals include details of the test, how the venue should be set up, steps required for preparation, how the test should be administered including the instructions to be read out to test takers, and how to store the filled papers. Scoring staff also gets hands-on training in marking Constructed Response Questions (CRQs). Practicing marking of CRQs can go a long way in improving the reliability of the test.

In addition to selection and preparation, a sufficient number of staff needs to be allocated to the test centers to ensure smooth implementation. For this, it is important that adequate funding is provided prior to test administration.

Test administration

The first step in test administration is selection of test centers. Typically schools are selected as test centers in Pakistan and elsewhere. Schools that will serve as test centers are usually identified with the assistance of local administrators. Selected schools are informed well in advance of test administration dates.

A set of SOPs is then prepared for the distribution, administration, collection, and return of test papers. The distribution and return of papers require strict adherence to SOPs and test papers are usually collected and returned on the day of the test. The SOPs are designed to ensure the security of test papers and to prevent possibilities of the use of unfair means. Invigilating staff is oriented to these SOPs.

During test administration, a monitoring team or teams observe the process in a sample of test centers. The monitoring is intended to further reinforce the secure administration of the test. The recommendations of the monitoring teams are used for improvement in adherence to the SOPs for subsequent rounds of test administration.

Test scoring

Scoring procedures vary with the type of items. Multiple Choice Questions (MCQs) are often scored by Optical Mark Recognition or Reader (OMR) software. The testing instruments designed to use OMR software for scoring require students to fill in bubble sheets that are scanned and marked by a computer. OMR software, however, cannot be used for items requiring open-ended responses. Although advanced algorithms have been developed for marking such responses, their use is still uncommon. Therefore, open-ended responses require the use of rubrics by scorers. Syndication is also used as a technique for such scoring. In syndication, each scorer marks only one question or a set of questions. Syndicated scoring ensures that students are not penalized or rewarded unfairly on the entire test due to scorer-related variation in scores.

Quality assurance in scoring is achieved through different methods. For scoring of MCQs, procedures for dealing with situations where students mark more than one option are handled by the computer program. In the case of open-ended questions, double scoring methods are utilized where a

second person independently scores the test. In practice, usually a certain percentage of test papers are rechecked by lead scorers. Another measure to ensure quality assurance in scoring is to monitor the number of papers or questions that are scored per session or day, as this has an impact on the quality of the scoring process.

PRIMARY AND ELEMENTARY LEVEL ASSESSMENTS

PUNJAB

The Punjab Examination Commission (PEC) exams are administered to approximately 3 million children in grades 5 and 8 annually. They are spread over a period of two weeks and are administered simultaneously across the province.

Test administration

Selection, preparation and allocation of staff

PEC exam supervisory staff comprises a resident inspector, supervisor, and invigilators, all of whom are exclusively practicing public school teachers and head teachers. Only the resident inspector is from the school where the exam center is established. No other supervisory staff is from the school in which the exam center is established. PEC provides the supervisory staff with its SOP manuals, which contain detailed written procedures for setting up and staffing exam centers, distributing and collecting papers, ensuring secrecy, and wrapping up exams. The superintendents also receive one-day training on their role. However, invigilators do not receive any training from superintendents.[4]

Supervisory staff absenteeism and shortages, particularly of invigilators, have been noted as persistent issues when it comes to the PEC exams.[5] According to the SOPs, there should be one invigilator for every 30 students; however, that is not always the case. The resident inspector is expected to monitor staff shortage and procure additional staff. PEC has recently instituted further steps to address this issue. In order to manage invigilators better, there is now an online human resource database. Efforts to counter invigilator absenteeism have included tying their payments to presence and increasing their remuneration.[6]

Administration procedures

For the PEC exams, the administrative structure is set up at the district, cluster, and exam center levels. The structure is as follows:

- District examination cell- set up in every district with a PEC focal person appointed to ensure coordination.
- Cluster and distribution center- based in a high school, the cluster center is headed by a Cluster-in-Charge. The cluster center is responsible for coordinating with and distributing exam papers to a set of exam centers.
- Exam center- headed by a resident inspector, usually the head teacher of the host school, who is responsible for ensuring facilities and space for the exams. In addition to the resident inspector, the center has a superintendent, in some cases a deputy superintendent, and a set of invigilators. These personnel are, as mentioned earlier, school teachers selected from schools other than the one in which the exam center is located.

It has been observed that schools designated as exam centers often lack minimum facilities, including the furniture needed for properly seating test takers. Often, students are seated close together making prevention of cheating extremely difficult.[7] Although lack of facilities continues to be a challenge, PEC has attempted to address the issue of cheating by delivering more than one version of the paper to each exam center so that children seated close together cannot copy from each other. Exam centers are also not always adequately allocated and sometimes can be too far from the place of residence of participating students and teachers.

There have been far fewer reported incidents of problems with the printing, delivery, and secrecy of exam papers for the PEC exams. There have been some exceptions, including the 2014 exam cycle when it was found that the security of exam papers had been compromised.[8] In 2015, there were reported problems with the printing and delivery of papers.[9] However, in 2016, papers were printed and delivered on time without any evidence of security compromises.[10]

There is an elaborate protocol in place for ensuring secrecy of exam papers and preventing cheating during exam administration. This entails the resident inspector collecting papers

on the day of the exam, superintendents opening sealed papers in the presence of two witnesses, and the resident inspector returning the filled papers on the day of the exam. It also entails, as mentioned previously, providing different versions of the paper to the same exam center. Studies have found that there was a high level of compliance with PEC SOPs particularly with regard to maintaining secrecy and following proper administration procedures.[11]

There is significant monitoring of the exam administration process at different levels. Two mobile inspectors are expected to visit the exam center on any given day. Third party validations have also been conducted in 2015 to determine the extent to which the exam center staff implemented the SOPs.[12]

Test scoring

Selection and preparation of staff

The scoring of the PEC exams is done in marking centers which consist of one head examiner and several examiners. In previous years, the PEC scoring process had been lacking. For instance, PEC did not have a sufficient number of examiners with the appropriate subject specializations.[13] In response, the number of examiners has been increased over the years and the head examiners are now subject specialists. Examiners are also now given training (using a cascade model) on utilization of the rubrics for scoring items in different subject domains.

Scoring procedures

The scoring process entails marking both the CRQs and MCQs by hand. Up until 2014, the MCQ portion of the PEC exam was marked through OMR software. This practice has been discontinued. An earlier study found that marking criteria were often not available or not used by the examiners due to lack of monitoring.[14] One examiner was expected to mark up to 100 papers per day. The examiners lacked subject specializations. Many resorted to over-marking. In nearly 50% of observed instances, marks were also allotted for wrong or irrelevant answers.[15]

PEC has instituted several changes in recent years to address the issues mentioned above. One examiner is expected to mark no more than 50 papers per day. Since 2015, the marking is syndicated to ensure consistency in the marking process. PEC has worked over the years to improve the quality of the rubrics, making detailed scoring schemes and ensuring the timely delivery of these to the marking centers.

The head examiner is responsible for overseeing the paper marking activity. This entails rechecking at least 10% of papers. PEC teams also monitor the marking centers. In future, PEC plans to conduct a marking analysis of papers to understand the impact of these improvements on the marking process.

SINDH

Sukkur Institute of Business Administration (IBA) has been administering the Standardized Achievement Test (SAT) to approximately 300,000 children in grades 5 and 8 annually since 2012. The test is administered sequentially by region over a maximum of ten days per region (there are five regions altogether).

Test administration

Selection, preparation and allocation of staff

Sukkur IBA test administration staff consists of taluka[16] coordinators and invigilators who are supervised by the SAT project coordinator and program officer. Invigilators are all graduates of Sukkur IBA; none of the invigilators are teachers or government officials. Invigilators are hired at the recommendation of the taluka coordinators. This large human resource is one of the key strengths of Sukkur IBA and the SAT administration is considered reliable largely because of the use of these graduates.

Sukkur IBA hires approximately 122 taluka coordinators at the taluka level that can cover three to four test centers per day over the course of the 10 days that the SAT is administered, and around 2,400 – 2,500 invigilators at the town level. This amounts to approximately one invigilator per 25- 35 students.

Taluka coordinators are given one-day training at the regional level. The selected invigilators also receive one-day training at the district level. Both trainings are conducted by the SAT project staff. Besides training them on their roles, the project staff also provides the participants with an overview of the purposes of the SAT and test administration mechanisms. Test administration staff is also given a comprehensive manual detailing administration procedures.

Administration procedures

The SAT project officers and taluka coordinators meet with the Supervisor Primary Education (SPE) and the Assistant District Officer Education (ADOE) to identify test centers (usually elementary and secondary schools) and feeder schools. If there is no elementary school in a 1.5 km radius of a primary school, the primary school serves as a test center. Once the list of center and feeder schools is finalized, the SAT project officers get these lists stamped by SPEs and ADOEs

or the Taluka Education Officers in every taluka.[17] There are approximately 15,000 test centers across Sindh where the SAT is administered.

Regional education workshops are conducted with the Regional Director Education, SPEs, ADOEs, and Taluka Education Officers. These workshops are used to explain the purpose of the SAT, logistical details, and the final schedule for the administration of the test. Sukkur IBA also sends unused booklets from previous years to head teachers so that they know what to expect in the SAT. This is useful, as a substantial proportion of these teachers do not have access to the internet and are unable to consult the online item bank. A few days before the test administration, taluka coordinators contact head teachers to provide the SAT schedule to them. For the third round of the SAT, Sukkur IBA also advertised the SAT in newspapers two months prior to the administration so that all stakeholders were aware and ready.

A courier service, TCS Pvt Limited, prints and dispatches pilot and actual test papers. On the day of the test, Sukkur IBA sends test booklets to the taluka coordinator in the morning via TCS. The taluka coordinator then distributes these among the test centers before 10 a.m. The taluka coordinator collects the test booklets from test centers after the test has been completed and sends these back to Sukkur IBA the same day through TCS to minimize chances of cheating.

Monitoring of test administration is done by Sukkur IBA, Reform Support Unit (RSU), and the district education offices. 22 senior project officers and managers from Sukkur IBA conduct the monitoring. In each district one senior project officer is accompanied by three or four monitoring team members. The team may include the SAT project coordinator, project officers, faculty members, alumni, and the more competent taluka coordinators. The monitors receive training from the SAT project office on the use of monitoring tools. The RSU also monitors administration of the SAT through its representatives at the district level, known as Local Support Units.[18] Although Sukkur IBA has reported difficulties working with Local Support Units' staff in some districts, details about the difficulties have not been provided.[19] Finally, Executive District Officers and District Education Officers also conduct random monitoring visits to test centers.

One set of challenges relates to low participation rates in the SAT. Children often do not show up on the day of the test in part due to lack of awareness about the SAT among students, parents, teachers, and even some education managers, and in part because of test centers being located far away. In many instances, head teachers are not informed about test schedules in a timely manner. For example, during the third round of the SAT in 2015, head teachers had not been informed in 80% of the cases that the test was due to take place. This resulted in only 55% of students taking the test.[20] The government is supposed to issue a letter about the test, but often head teachers say they have received no prior notification. The RSU has conducted some motivational workshops in the last two years to raise awareness and improve communication regarding the purpose of the SAT so that more students take the test.

Another set of challenges is associated with resistance from head teachers and teachers' associations in allowing students to be assessed. However, after several successive rounds of the SAT, the level of resistance has decreased. Sukkur IBA has also taken measures to ease head teachers' concerns by informing them that the test is not high-stakes. Nevertheless, the teachers' associations at both the primary and secondary levels sometimes shut down test centers. When this happens, Sukkur IBA has to involve the district administration for help.

Another challenge is cheating and interference by local officials in the test administration process. For instance, a taluka coordinator once reported that the test center had been taken over by a local administration official where students were being helped in answering questions on the test. In this instance, the coordinator was asked to tag these booklets and Sukkur IBA did not include them in the marking and analysis of test results. Sukkur IBA reports these incidents to the RSU; however, often action is not taken.[21]

Finally another challenge is that delays in the test cycle have an impact on the number of students available to take the test and ultimately the validity of the results. In many instances, RSU does not release funds to Sukkur IBA according to schedule. As a result, this throws off the entire test cycle and instead of the test being administered in January/February before final exams, it takes place in March when fifth graders are now in grade 6. Between grades 5 and 6, there is substantial reduction in the size of the population taking the test due to the high dropout rate. It can also be argued that testing students in grade 6 biases the results upwards since students who have now dropped out were more likely to have been weak performers.

Test scoring

Selection and preparation of staff

Each scoring team consists of e-markers[22], reviewers, and a super reviewer. For the purposes of test scoring, Sukkur IBA enlists about 200 e-markers out of which it selects 100. These e-markers attend a three-day training workshop in which the SAT project team conducts a marking exercise followed by a session to review issues and problems that had arisen in it.

Scoring procedures

Every booklet is assigned a unique barcode to keep student information anonymous. MCQs are automatically marked using OMR software. This has been the practice since the beginning, with the exception of the first SAT which was manually marked.

CRQs are forwarded to e-markers. Each e-marker is assigned one to two questions and they work in groups of five to eight. E-markers are guided by a scoring scheme which contains an analytical rubric as well as an example of an answer for every CRQ that appears on screen. From 2016, all items in the SAT, including CRQs will be marked via OMR software using algorithms based on rubrics. These will be constantly redefined based on feedback. During the e-marking process of the SAT, 15% of questions are rechecked and verified by assessment experts and feedback is given to e-markers. A further 10% of items are rechecked by a super reviewer.[23]

KHYBER PAKHTUNKHWA

The Khyber Pakhtunkhwa (KP) government has begun to administer large-scale standardized assessments to all students in grade 5 public sector schools. In 2015, approximately 300,000 students were tested in English, Urdu, math, and science.[24] In 2016, approximately the same number was tested in six subjects- English, Urdu, math, science, social studies, and Islamiat.[25] Multiple versions of test papers were prepared. Apart from the Urdu test, tests were administered in the English language.[26] Given that large-scale assessments are relatively new in KP, limited information about its administration and scoring is available.

For sample-based assessments, 2015 was the first year in which the Provincial Education Assessment Center (PEAC) conducted an assessment since 2008. It tested approximately 4,961 students in grade 2 (who had just entered grade 3) in approximately 350 schools in Urdu, English, and math.

Test administration

In KP, the government has relied on the existing capacity of the Boards of Intermediate and Secondary Education (BISE) to administer the large-scale assessment. All eight BISE are involved in the administration of the grade 5 assessment. For this administration, the BISE are using their existing setup that they mobilize for secondary and higher secondary exams. Province-wide administration was handed over to the BISE because the government could not find a credible private sector organization to deliver these tests; an attempt to do so in 2015 was not successful. No other institution is perceived to have the capacity with approximately 12,000 personnel needed in the 3,000 test centers across the province. The assessment for grade 5 was successfully administered by the BISE in April 2016. No aspect of the 2016 test administration was outsourced. For the time being, scoring and other quality concerns are also being handled by the BISE.

The sample-based assessment was conducted by PEAC in 2015. Moving forward, there is some ambiguity regarding which aspects of the 2016 sample-based assessment will be outsourced to a third party, if at all.

Test scoring

For large-scale assessments at the grade 5 level, the test scoring has been done differently by different boards. Except for the Peshawar BISE, nearly all the BISE are resorting to manual scoring done by appointed scorers. No scoring rubrics have been developed so far. However, for the 2017 assessment, the Directorate of Curriculum and Teacher Education and the BISE teams are working together on developing such rubrics. The marking of the PEAC sample-based assessments has been done manually so far.

SECONDARY AND HIGHER SECONDARY LEVEL EXAMINATIONS

BOARDS OF INTERMEDIATE AND SECONDARY EDUCATION

The BISE administer Secondary School Certificate examinations in grades 9 and 10, and Higher Secondary School Certificate examinations in grades 11 and 12 every year in multiple subjects. They usually administer two papers per day from March to July, with the exact time varying across the different BISE. The number of students being tested also varies according to the population of the area that falls under a particular board's jurisdiction. For example, in boards such as Karachi Board of Secondary Education (BSE), approximately 170,000 students appeared in each of the grade 9 and 10 exams in 2016,[27] whereas other boards such as the Sahiwal BISE caters to fewer students, approximately 74,000 for grade 9 alone in 2015.[28] The number of students appearing for Higher Secondary School Certificate exams tends to be lower in comparison to those appearing for Secondary School Certificate exams.

Test administration

Selection, preparation and allocation of staff

Exam administration staff consists of superintendents, invigilators, and inspectors, all of whom are usually public school teachers. They are normally nominated by schools and education offices, and the list is finalized by the secret branch of the BISE. Exam center staff is allocated centrally and dispatched to a district different from the one they are from. One invigilator is made responsible for 40 students. The BISE Calendar provides detailed procedures for exam administration. However, there is no evidence of any initiative to train staff in exam administration, particularly on detection and prevention of cheating.

Administration procedures

The BISE issue guidelines on the establishment of centers, student registration, distribution and collection of papers, and prevention of unfair means. The BISE approach the district administration for making security arrangements at the exam centers. Exam materials are distributed in different ways. In some cases, the board distributes the papers directly (e.g. the Karachi BSE) and in other cases, the superintendents collect them from a local bank. After the exams have been administered, the superintendents are responsible for returning the filled papers to the bank within a stipulated time frame (e.g. in the case of KP, the filled papers must be returned within an hour of the exam being completed). The BISE collect the filled papers and send them over to the marking section of the boards.

The use of unfair means throughout the exam cycle remains a major issue in the BISE examinations. There are opportunities for resorting to unfair means at several stages: influence in choice of centers, exam staff and inspectors; possibilities of paper leakage; and cheating during examinations which includes support from invigilators in the form of answers, extra time, use of notes by candidates, and candidates helping each other.

Notwithstanding the above, the governments and the BISE have taken several administrative measures to curb these behaviors. For example, in Punjab, the BISE take action against individuals involved in the use of unfair means in accordance with the Punjab Malpractice Act of 1950. In addition, special monitoring bodies have been established - the Examination Monitoring Cell in Punjab and the Governor appointed Special Vigilance Teams in Sindh - to supplement measures adopted by the BISE. The BISE have also taken steps to avoid leakage of exam papers by collecting and depositing the papers on the day of the exam at banks. Furthermore, different versions of the papers are distributed. Governments have begun to use a scanned picture on the registration form in order to prevent identity impersonation. The exam booklets are also codified in order to ensure anonymity of the student. Despite these efforts, cheating remains firmly entrenched in the system. Lack of accountability, particularly with regard to examination staff, is considered the main reason behind the persistence of unfair activities.

Test scoring

Selection and preparation of staff

The scorers consist of head examiners and examiners. No training is provided to them on marking the exams. They are, however, given guidelines for the scoring of papers with the assumption that they will use them correctly without any further guidance.

Scoring procedures

Scoring in the BISE, with the exception of Punjab and the Peshawar BISE, is done by hand. Paper setters provide a marking key and head examiners provide broad guidelines to the examiners on how to score. However, for the most part, examiners rely on their experience and subject expertise to score papers. The number of papers that the examiners are expected to score each day varies from 25 to 50. In practice, however, they end up scoring up to a 100 papers a day. The head examiners are expected to review a certain percentage of papers, varying from 5% to 10%. However, this does not always happen. There is also a combination of scoring at the center and scoring at home. There are concerns that the scoring process is inadequately supervised and monitored.

Some efforts to improve this process have been made in recent years. In Peshawar, attempts have been made to develop standardized marking keys for each subject by the examination staff. The Peshawar BISE has also added a super checker, usually a seasoned professor, who is responsible for reviewing the marking key and rechecking a percentage of best and worst papers. In Sindh, there is an effort to ensure that the paper marker is from the same language background as the paper they are checking. In Balochistan, Closed Circuit Television (CCTV) cameras have been installed to monitor marking at the centers.

In Punjab and in the Peshawar BISE, efforts have been made to improve quality of marking by: (1) developing machine readable answer sheets that enables checking of MCQs by OMR software, thus reducing opportunities for the use of unfair means and examiner burden; (2) swapping papers between boards to reduce opportunities for unfair means; and (3) using syndicated marking in which one examiner marks only one question, in an effort to reduce marker bias.

The success of any of these efforts is not known clearly. Despite these measures, stakeholders as well as several research reports consider scoring as one of the weakest aspects of the examination process at the BISE.[29]

AGA KHAN UNIVERSITY-EXAMINATION BOARD

Test administration

Selection, preparation and allocation of staff

Supervisors and Invigilators are hired by Aga Khan University- Examination Board (AKU-EB) on a part-time basis and training sessions for these supervisors are conducted. Case studies from past years are discussed to form a consensus around the best ways to address instances of potential cheating. Emphasis is placed on ensuring that any punitive actions do not infringe on the dignity of the candidates.

Administration procedures

The operations department is responsible for making the logistical and supervisory arrangements necessary for exam administration. While creating the timetable, AKU-EB makes sure that there is a sizeable gap between subjects generally regarded as tough to aid students in their preparation schedules, thus ensuring that the timetable is student-friendly. The operations team provides sound equipment to each examination center to play recordings during the listening component of the language exams. They also make sure that students with disabilities are properly accommodated. Students with dyslexia, autism, or dystonia are provided extra time and a comfortable seating arrangement. Papers with increased font size are arranged for the visually impaired, and students with mild aural disabilities are facilitated during the listening component of the exam.

The centers have two checkpoints before students enter the exam room where they have to deposit any technological devices, such as mobiles or smart watches that could potentially be used as an aid during the exam. If caught cheating, students are allowed to finish the examination and sign a form at the end as confirmation that their case was flagged by staff. All such malpractice cases are reviewed by a panel within AKU-EB. The panel considers all the contextual information and any extenuating circumstances while making a decision; an appeals process is available for candidates not satisfied with the decision.[30] The panel reviews unique cases arising from unforeseen circumstances every year. In 2015, a CCTV camera was installed in one of the centers. In 2016, over 75% of the exam centers across Pakistan were monitored through CCTV cameras. Each exam center has multiple monitoring cells. Moreover, overall monitoring is done at the AKU-EB head office in Karachi. The board plans to cover all exam centers with CCTVs in 2017.

Test Scoring

Selection and preparation of staff

Scoring staff consists of e-markers, senior e-markers, and supervisors. Inexperienced teachers from AKU-EB affiliated schools are encouraged to apply as e-markers as the process of participating in the marking procedures is considered to be a form of professional development. Supervisors, including senior e-markers familiar with the process, ensure these e-markers are provided with the requisite support.

Scoring procedures

When the filled papers arrive for marking at AKU-EB, the front cover is scanned and cut. Only the computer servers can retrieve the identity of the student. The e-markers are unaware of the identity of students whose papers are being scored. The electronic copies are split into their component sub-sections and items. MCQs are marked by OMR software, and responses to CRQs and Extended Response Questions (ERQs) are fed to e-markers by the server.

The staff responsible for scanning the MCQ sheets glance over them while scanning to ensure that options are filled properly, and flag cases where the OMR software may not be able to detect the selected answer (e.g. where the student has filled an option improperly or left room for ambiguity). While MCQs are electronically scanned and scored, the MCQ marking keys for each of the different versions of the exam paper are reviewed to ensure there are no mix-ups or problems with sequencing.

Scoring of CRQs and ERQs is syndicated to ensure scoring consistency across different papers. E-markers work in groups of four with each group being supervised by a senior e-marker, who in turn is overseen by a supervisor. During marking of the CRQs and ERQs, e-markers have the option of flagging responses that are unclear. If students compose a response using a method not accounted for in the rubric, the e-marker can select the option "needs to be discussed further" so that more experienced staff members can deliberate on how to score the flagged question.

The senior e-marker can view the scoring patterns of the whole group and intervene if needed. The senior e-marker is able to view information about the time taken by a particular marker to score questions and can step in to guide an e-marker who takes a shorter or longer time than recommended for a specific question or if there are inconsistent marking patterns over time. The entire marking process is overseen by the AKU-EB Director and Associate Director of Assessments along with supervisors, and backend support is provided by the operations team. Content experts also review the scored papers to ensure that they have been marked properly.

AKU-EB has several layers of quality assurance for the exam scoring process. The scoring of papers is also analyzed to ensure reliability and fairness of results. AKU-EB performs an analysis of the marking of all versions. The two quality indicators in the analysis are the difficulty index and the discrimination index. Each item from all exam papers is reviewed with the teacher training unit at AKU-EB that developed the syllabi and the items. The team assesses whether students performed as expected and judge if an unexpected score was the result of the student not being acquainted with a difficult topic or if, in the case of MCQs, available options were too similar to distinguish the correct response. Items that are judged to be constructed improperly are disqualified. If it becomes apparent that two options for the same item can be considered correct, full marks are awarded if either option is selected. If three options are deemed correct, full marks are awarded regardless of the selected option.

As an additional check on the rubrics and scoring schemes for CRQs and ERQs, a staff member selects 30 students representative of all geographical regions of the country and all performance levels as indicated by MCQ scores. Two senior e-markers are given the scoring rubrics to mark the same 30 papers independently and a report is generated from the results. The e-markers also give qualitative feedback on the scoring rubrics. The report is discussed by content experts who compare scores given by the e-markers to ensure standardization across scores. In case of any discrepancy arising from both e-markers having marked the same paper differently, the scoring rubrics are revised for greater clarity and precision, and the whole process is repeated until discrepancies no longer arise. Content experts along with these senior e-markers subsequently train the team of e-markers on appropriate use of the rubrics.

Additional review processes include: (1) scores obtained in MCQs, CRQs, and ERQs in a particular subject by the same student are compared. The expectation, borne out by research,[31] is that a student who performs well in the CRQ portion of a subject should perform well in the MCQ section as well for that subject. If there is significant variation, the filled MCQ sheet is reviewed to check that the correct barcode was scanned and the OMR software detected the selected options properly; and (2) item writers, content experts, and the teacher training unit collectively review student responses for the purpose of informing and positively enhancing assessment design practices for all staff involved.

CONCLUSION

Across the cases discussed above, there is slightly more alignment with best practices when it comes to assessment implementation (refer to Table 5.1).

In most cases, assessment agencies make use of existing school teachers to administer and score tests, with the exception of Sukkur IBA who hire alumni of the university to administer the SAT. Using teachers is a common practice in many countries but they are usually not practicing teachers or are from non-participating schools. Given the high-stakes nature of many of the tests, using practicing teachers whose schools are also being tested can prove problematic as they have a vested interest in the outcomes of the assessment. There is a need to rethink the selection criteria for the implementation staff in several assessment systems along with greater monitoring of the implementation process.

Training of the administrative and scoring staff as well as provision of manuals appears to be the norm amongst PEC, Sukkur IBA, and AKU-EB while the BISE provide no such training for any of their staff. AKU-EB appears to be the only system which conducts extensive training for its administrative staff on how to handle instances of cheating.

The administration of these large-scale assessments and examinations is an arduous task. All assessment agencies appear to have sufficient mechanisms in place for distribution and collection of papers and allocation of test and exam centers. In the case of exams administered by PEC and the BISE, controlling cheating is a major issue due to the high stakes of the exams. While PEC has managed to prevent instances of cheating better, the BISE still struggle. Sukkur IBA faces fewer instances of cheating given the low stakes of the SAT. On the other hand, it faces issues related to non-participation. There is no publicly available information or media reports on cheating in AKU-EB exams. It is clear that AKU-EB has elaborate procedures for dealing with instances of cheating and also far fewer numbers of students to deal with unlike the other BISE.

The test scoring process has traditionally been one of the weaker aspects of assessment systems in Pakistan. This is particularly so in the case of the BISE exams and little has been done to improve these practices. Amongst the primary and elementary level assessments, there appears to be congruence with best practices. Marking of the MCQs using OMR software has become the norm for the most part, even some of the BISE use this practice. PEC's departure from this practice, particularly given the scale of the exam, appears problematic. For marking of CRQs, scorers are now provided with rubrics and detailed scoring schemes with the exception of the BISE which just follow general guidelines. The processes for quality assurance are in place, which entail rechecking a certain percentage of papers.

Table 5.1 Assessment implementation practices in Pakistan

	PEC	SAT	KP G5	BISE	AKU-EB
Training administrative and scoring staff	✓	✓	Insufficient information	×	✓
Providing manuals with SOPs	✓	✓	Insufficient information	✓	✓
Using OMR software to score MCQs	×	✓	Practice varies	Practice varies	✓
Using rubrics to score CRQs	✓	✓	×	×	✓
Monitoring administration and scoring	✓	✓	✓	✓	✓

[1] American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999)
[2] Greaney & Kellaghan (2012)
[3] Greaney & Kellaghan (2008)
[4] Exam center staff interviews (15 March 2016)
[5] Ibid.; SAHE (2011)
[6] "5th, middle class exams" (2016)
[7] SAHE (2011)
[8] "CM orders probe" (2014); "8th class math" (2014)
[9] Exam center staff interviews (15 March 2016)
[10] Ibid.
[11] SAHE (2011)
[12] Exam center staff interviews (15 March 2016)
[13] SAHE (2011)
[14] Ibid.
[15] Ibid.
[16] Tehsils, which are administrative sub-units of a district, are referred to as talukas in Sindh
[17] RSU & Sukkur IBA (2015)
[18] Local Support Units are a part of RSU and were set up in 2013 – 2014 as per SERP II. The unit comprises of a District Coordinator - a Grade 17 officer selected from within the Sindh Government - and a Local Support Consultant - recruited from the private market as part of a competitive hiring process.
[19] RSU & Sukkur IBA (2015)
[20] Ibid.
[21] Ibid.
[22] The titles assigned to the staff associated with different assessments systems may vary even if their responsibilities are largely similar, e.g. AKU-EB and IBA-Sukkur both refer to those scoring papers as e-markers, as the filled booklets are checked electronically, while PEC and BISE refer to their checkers as examiners.
[23] RSU & Sukkur IBA (2015)
[24] "5th grade assessment exam for elementary" (2015)
[25] Directorate of Curriculum and Teachers Education (DCTE) (2015)
[26] "Grade 5 exams: Educationists demand Pashto" (2016)
[27] "Across the metropolis" (2016)
[28] "9th Class Annual Exam results" (2015)
[29] Shah & Afzaal 2004; Bethell, Dar, & Crighton (1995), as cited in Christie & Afzaal (2005)
[30] AKU-EB staff interviews (4 December 2015)
[31] Mujeeb, Pardeshi & Ghongane (2010)

06

DISSEMINATION AND USE OF ASSESSMENT RESULTS

INTRODUCTION

The discussion in the preceding chapters focused on refinements, opportunities, and challenges pertaining to assessment design and implementation practices in Pakistan. This chapter will focus on the development of assessment results, their communication, and use. Appropriate analysis, meaningful interpretation, and timely dissemination of assessment results are essential to driving improvement in the education system. It is imperative that the quality of the education system be improved continually through an iterative process involving conduct, communication, and use of assessments.

This chapter begins by delineating best practices in communication and use of assessments to drive improvement in the education system. This discussion is followed by a review of alignment, or lack thereof, of practices in Pakistan. Each case of assessment in Pakistan includes a description of what and how results are developed, how they are communicated to different stakeholders, and how, if at all, assessment results are used.

BEST PRACTICE[1]

Developing results

Analysis of the scored data is typically conducted by a data analyst or statistician who works closely with the test design team. A data analyst's or statistician's work entails analyzing data generated during the test pilot to assist with selecting items and analyzing data generated when the actual test has been conducted to produce final results.

Assessment reports usually present results in the form of mean scores. In addition, analysis may be tabulated in different ways that include:

- Single index versus multiple indexes; either one score per subject can be tabulated or the scores can be tabulated according to the cognitive domain. For example, in math, this would entail providing separate scores for computation, measurement, problem solving, and so on rather than just one math score. Providing multiple indexes allows for reporting a more holistic picture of student achievement with greater opportunity for use.

- Scores versus proficiency levels;[2] scores by themselves provide a limited amount of information about student performance, while proficiency levels indicate the degree to which students have mastered the material, and, hence, they are able to provide more information on student strengths and weaknesses. Proficiency levels have become increasingly popular in reporting assessment results. Each proficiency level is usually accompanied by a range of scores associated with it and a subject specific description of what each level means in terms of student ability (refer to Tables 6.1 and 6.2). The nature of proficiency levels varies for different assessments depending on the standards on which they are based. This impacts how proficiency levels are determined (i.e. what the cut scores are) and how they are inevitably reported.

Table 6.1 Percentage of students reaching international benchmarks mathematics TIMSS grade 4
Source: Adapted from TIMSS (2003), p. 63

	International Benchmark			
Countries	**Advanced (625)**	**High (550)**	**Intermediate (475)**	**Low (400)**
Singapore	38	73	91	97
Hong Kong, SAR	22	67	94	99
Japan	21	60	89	98
Chinese Taipei	16	61	92	99
England	14	43	75	93
Russian Federation	11	41	76	95

Table 6.2 International benchmarks of mathematics achievement TIMSS grade 4
Source: Adapted from TIMSS (2003), p. 63

Advanced International Benchmark – Scaled score 625
Students can apply their understanding and knowledge in a wide variety of relatively complex situations. They demonstrate a developing understanding of fractions and decimals and the relationship between them. They can select appropriate information to solve multi-step word problems involving proportions. They can formulate or select a rule for a relationship. They show understanding of area and can use measurement concepts to solve a variety of problems. They show some understanding of rotation. They can organize, interpret, and represent data to solve problems.
High International Benchmark – Scaled score 550
Students can apply their knowledge and understanding to solve problems. Students can solve multi- step word problems involving addition, multiplication, and division. They can use their understanding of place value and simple fractions to solve problems. They can identify a number sentence that represents situations. Students show understanding of three-dimensional objects, how shapes can make other shapes, and simple transformation in a plane. They demonstrate a variety of measurement skills and can interpret and use data in tables and graphs to solve problems.
Intermediate International Benchmark – Scaled score 475
Students can apply basic mathematical knowledge in straightforward situations. They can read, interpret, and use different representations of numbers. They can perform operations with three- and four-digit numbers and decimals. They can extend simple patterns. They are familiar with a range of two-dimensional shapes and read and interpret different representations of the same data.
Low International Benchmark – Scaled score 400
Students have some basic mathematical knowledge. Students demonstrate an understanding of whole numbers and can do simple computations with them. They demonstrate familiarity with the basic properties of triangles and rectangles. They can read information from simple bar graphs.

Disseminating results

Assessment results reports must fulfill two conditions. First, they must conform to the assessment framework and second, they must be accessible to a wide range of stakeholders who can potentially benefit from them. Results can be communicated in different ways to different stakeholders. Accordingly, development of a results report and its dissemination requires diverse and high level input. Apart from the main report, assessment agencies produce briefings for ministers or senior policy personnel which focus on key findings and issues along with recommendations; non-technical summary reports that target teachers and the wider population; technical and thematic reports for the research community; and press briefings and media reports.

In many instances, assessment agencies do not produce and communicate results effectively. The use of assessment results as feedback into the system for its improvement remains an unsolved problem. On the other hand, assessment results are often used to hold teachers accountable for students' performance. Such use of assessment results is controversial since findings of studies investigating the link between teacher performance and student learning gains remain inconclusive.

Using results

Assessment results can be used in the following ways: (1) to describe and identify gaps in student achievement; (2) to monitor and evaluate the effectiveness of policies and interventions in the education sector; (3) to address deficiencies identified by the assessment through formulating policies, revising curricula and textbooks, informing teaching practice and teacher professional development; (4) to hold districts, schools, and teachers accountable for the results produced; and (5) to certify completion of a cycle of schooling, promote to the next level, or select individuals for limited opportunities.

A review of assessments in the Asia-Pacific region shows that assessment results are used most frequently to inform system level policies on assessment and resource allocation policies targeting in-service professional development programs and instructional materials. Assessment results are used least for informing in-classroom teaching and learning practices and policies.[3]

There are several factors that affect the use of assessment results:[4]

- Level of integration with the policy process; to begin with, legally mandated assessments are more likely to be integrated in the policy process. Other factors also contribute, such as whether the assessment is perceived as a stand-alone activity or integrated with other educational activities, whether there has been adequate involvement of stakeholders in design and implementation of an

assessment, or whether there are actually plans to devise policy or school level actions based on assessment data.

- Quality of the assessment system; results will be used only if stakeholders have confidence in the validity and reliability of results. Lack of confidence in the findings of assessments can be an issue due to the technical quality of the assessments as well as the quality of implementation.

- Effectiveness of the communication strategy; assessment results need to be effectively communicated to ultimate users of these results. A communication strategy that ensures rapid communication of results, in the form of accessible reports, to all stakeholders is essential.

Finally there are certain limitations in using assessment results for accountability.[5] To begin with, using outcomes as the sole basis of accountability ignores the fact that student performance is influenced by more variables than just teacher's performance. These could be: (a) characteristics of students; (b) family related influences or conditions in which students live, including community resources and support; (c) education policies and resources and support, including curricula and teacher preparation, that are provided by the relevant public authorities; and (d) school conditions and resources, including school governance and management.

Secondly, assessments measure cognitive abilities, whereas schooling has many other non-cognitive purposes as well, for example social and moral development. Therefore, a reliance on cognitive measures is not sufficient to assess a teacher's effect on student development.

Thirdly, providing incentives or sanctions based on student performance often leads to negative consequences such as teaching to the test, emphasizing certain skills such as rote memorization, drilling of knowledge that results in a passive approach to learning, an increase in time spent developing test taking strategies, and/or even excluding low performing students from participating in assessments.

Many assessment agencies recognize the role of such factors and present assessment results according to the socioeconomic status of students or schools. The purpose of highlighting these additional variables is to caution against the sole use of results of student assessment to penalize or reward teachers.

PRIMARY AND ELEMENTARY LEVEL ASSESSMENTS

PUNJAB

Analysis and reporting of exam results has been one of the weaker areas of the Punjab Examination Commission's (PEC) work. PEC lacks the staff to conduct data analysis and to develop reports. Lack of capacity also limits PEC's ability to produce results and reports in a timely manner and for a wider range of audiences. This is discussed in some detail below.

Developing results

PEC develops aggregate and subject-wise mean scores based on the analyzed data. It disaggregates scores by gender, district, and school type, language of exam, question type, cognitive domain, topics, and Student Learning Outcomes (SLOs). Currently, results are provided to all stakeholders in the form of mean scores only. In the future, PEC is considering reporting results in terms of three different performance or proficiency levels, which will provide more detail on student performance.

Disseminating results

Currently, the focus of PEC's communication has been on producing results for individual students and districts. PEC develops both aggregate and subject-wise results and makes them available through printed gazettes, its website, and individual student report cards. It also develops a district level analysis for district staff in which scores are provided according to tehsil, gender, location, sector (i.e. public or private), and so on, but this is not widely shared in the form of a public report.

For the provincial level analysis, PEC develops one detailed report on results. Till now, such reports have been produced irregularly due to scarcity of technical staff. For example, reports were produced from 2008 - 2010 after which no report was produced till 2015. Moreover, for the 2015 cycle,

the analysis was conducted with support from a Department for International Development (DFID) supported technical assistance organization. In future, however, it is expected that PEC will independently conduct a regular analysis of exam data. The 2015 report on PEC exam results contains disaggregated mean achievement scores. Most importantly, the 2015 report introduced mean scores by cognitive domain, topics, and SLOs (refer to Table 6.3). This makes the report more relevant for provincial government officials as well as teacher professional development and teaching personnel.

Table 6.3 Example of how PEC examination results were reported in 2015
Source: Adapted from PEC (2015), p. 45

SLO	Performance on Application Based Questions	Performance on Knowledge Based Questions	Performance on Understanding Based Questions	Overall Performance	Number of Questions
MATHEMATICS	60.16%	54.99%	64.69%	61.28%	
DECIMALS AND PERCENTAGES	57.50%	38.73%	61.73%	56.29%	30
Add and subtract decimals			70.55%	70.55%	7
Divide a decimal with a whole number			59.65%	59.65%	4
Divide decimals by 10, 100 and 1000			37.21%	37.21%	1
Multiply a decimal by a decimal (in the same way as for whole numbers and then put in the decimal point accordingly)			50.06%	50.06%	1
Multiply a decimal by tenths and hundredths only		39.05%		39.05%	1
Multiply a decimal with a whole number			40.64%	40.64%	2
Multiply decimals by 10, 100 and 1000			70.06%	70.06%	3
Recognize like and unlike decimals	58.49%	40.93%		46.78%	3
Round off decimals up to specified number of decimal place		37.16%		37.16%	3
Solve real life problems involving decimals	57.30%			57.30%	5

The reports serve as the primary means for disseminating results to inform teaching, teacher training, and textbook development policies and practices. PEC additionally shares raw data that has been generated with the Directorate of Staff Development (DSD) in Punjab.[6] For the 2015 cycle, results were shared with key stakeholders and the heads of various education departments in the government. For future exam cycles, PEC plans on disseminating and discussing results through panels, conferences, and posters.

Using results

The PEC exam results are used to promote students. They are also used to determine student eligibility for various public scholarships such as those managed by the Punjab Educational Endowment Fund.

At the provincial and district levels, it is not clear how the results are used, if at all. On several occasions, the Punjab Government has revealed its plans to use exam results to incentivize teachers and hold them accountable (i.e. through tying promotions and salaries to results). In 2011, it introduced a High Achievers Program in which teachers and head teachers received incentives in the top 20% of schools; PEC exam results accounted for 70% of the performance evaluation of schools and teachers.[7] In 2013, PEC results were used to issue show cause notice to approximately 6,000 teachers whose students had performed poorly, that is, if their classes had a pass rate of 25% or less.[8] This happened again in 2014. There was a great deal of outcry over this policy, with teacher unions resisting it on the grounds that many other factors influenced student performance.

Considering that PEC exam data and results are shared with the DSD, they should potentially inform setting of priorities and planning for teachers' professional development in the Punjab province. Some DSD officials have, however, expressed reservations about the use of PEC data due to mixed perceptions about the quality of exam results. In addition, as mentioned earlier in Chapter 3, DSD has its own grade-wise monthly assessment system, which it uses to review its progress and to inform its professional development activities.

SINDH

Developing results

The analysis of test results is undertaken by the Sukkur Institute of Business Administration (IBA) staff itself. The statistics generated by the software used are descriptive in nature. As per the Terms of Reference, Reform Support Unit (RSU) only requires a descriptive analysis. Sukkur IBA also carries out a content strand analysis at the district, regional, and provincial levels for both grades 5 and 8 (refer to Table 6.4). As part of the analysis of results, the mean score is disaggregated by gender, school, district, and tehsil. The report does not provide any analysis of the predictive factors such as head teacher or teacher background.

Table 6.4 Example of how SAT results were reported in 2015
Source: Adapted from RSU & Sukkur IBA (2015), p. 43

Subject	Content Strand	Content Strand Average (%)	Subject & Overall Average (%)	Standard Deviation
Language	Reading	54.16	32.81	18.6
	Writing	11.47		
Math	Number & Operation	18.70	18.22	12.78
	Measurement	37.74		
	Geometry	14.65		
	Information Handling	11.56		
Science	Life Science	14.76	15.26	11.04
	Physical Science	14.49		
	Earth & Space Science	28.46		
Overall Students' Score (%)			**22.10**	**11.79**

Disseminating results

Sukkur IBA sends individual student report cards for grade 8 students to secondary schools. It encourages principals to organize parent-teacher workshops to discuss results with parents. However, there is no evidence yet about the use and effectiveness of this exercise. With regard to grade 5, disseminating report cards is difficult as many of the primary schools in Sindh do not have a postal address. In such cases, Sukkur IBA disseminates the report cards through the Assistant District Officer Education and Taluka Education Officers who then pass these along to head teachers who give these to parents.[9] Sukkur IBA also prepares school level reports that are shared with head teachers of schools via Taluka Education Officers. Results are also made available online on the website set up for the Standardized Achievement Test (SAT) by Sukkur IBA. In 2015, the website received over 10,000 hits.[10]

Sukkur IBA prepares one comprehensive report each year on test results. These are shared with the World Bank, RSU, provincial ministers, education officers, school administrators, and other stakeholders across the province and Pakistan. The RSU intends to pursue dissemination to a wide variety of stakeholders with greater seriousness in future.[11]

Along with communicating results, Sukkur IBA also proposes recommendations that can be adopted by the Sindh Government. Three broad categories of recommendations are provided on capacity building of teachers, on improving pre-service teacher education, and on research needs (refer to Box 6.1).

Using results

One of the challenges faced once SAT results are disseminated is whether and how these results will be used by stakeholders to improve student learning. Although Sukkur IBA produces recommendations, taking actionable steps remains the responsibility of the Sindh Government. Thus far, it appears that result use is very limited: districts lack the capacity to interpret results and there is limited interest from institutions such as the Bureau of Curriculum and Provincial Institute for Teacher Education (PITE).[12]

RSU is aware of the importance of using results to inform curricula and teacher professional development. District-wise groups have been set up to work on reviewing the curricula on the basis of SAT results data. In future, RSU will inform PITE about which trainings to deliver to teachers to improve student learning as well.

There appears to be interest amongst the Sindh Government to use the test results for teacher accountability. Thus far, Sukkur IBA and other stakeholders have advised against using results for this purpose, given that student performance cannot be attributed solely to teacher performance and needs to be evaluated in the context of prevalent socioeconomic conditions as well. The results of the previous rounds of the SAT have been used by RSU to incentivize teachers. However, this has not been done on a regular basis. RSU plans to incentivize students and high performing schools in future with prizes and certificates signed by the Secretary Education.

Finally, RSU also wants to make the SAT an exam (changing the purpose, therefore, of this assessment) and making it more high-stakes by tying children's promotions to it. One of the challenges with regard to the SAT is striking a balance between the SAT achieving its objectives while remaining a low-stakes test. The SAT needs to be taken seriously by stakeholders if it is to play a role in reforming the education sector. However, the SAT team will need to be creative about how to get stakeholders to be serious without making the

BOX 6.1 SAMPLE SUKKUR IBA RECOMMENDATIONS FOR IMPROVING TEACHING

"Sharing effective resources on teaching and learning of geometry with schools and training teachers to enhance their ability to teach this least- focused content strand effectively: Moreover, it seems that geometry is the most neglected area in teaching and learning of math as students' scores turned out to be severely low in both grades. Specific trainings should be conducted on content knowledge and pedagogical content knowledge for teaching geometry. Also several kinds of reading resources should be shared with teachers. UC-based and/or taluka-based interaction (workshops, reflective meetings, demonstration of effective lessons followed by critical and reflective discussions) among math teachers focused on teaching and learning geometry also will help in enhancing teachers capacity to teach geometry effectively.

Developing teachers science conceptual knowledge and science inquiry skills by engaging them in effective training and by sharing effective science reading as well as science teaching and learning materials: Students performance was found lowest in science and almost in all content strands; specifically quite low in physical science. It is generally observed that most of the teachers engage students in rote learning of science materials instead of developing their conceptual understanding and inquiry and problem solving skills. Teachers should be given effective reading materials on teaching and learning science focusing on conceptual understanding and inquiry skills and attitude. There are many effective resources available online; teachers should be engaged in utilizing those resources."

Source: RSU & Sukkur IBA (2015), p. 110-111

test high-stakes and distorting incentives.

As discussed in Chapter 2, RSU has established a taskforce with representation from various stakeholders and institutions who will now use and evaluate test data from the first three rounds of the SAT to supposedly launch in-depth documentation projects and interventions on the basis of these results. The relevant government authority will also focus on math and science expertise when recruiting teachers, as a result of findings from SAT results data.

According to Sukkur IBA, there is greater need to conduct further analysis on the existing test results data to design interventions to address the abysmally low student performance across Sindh. In light of this, Sukkur IBA is planning on undertaking case study based research to look into factors that contribute to high and low performance in schools. Sukkur IBA is also interested in focusing on interventions, such as developing teaching material and organizing training sessions for teachers, to improve student knowledge in math and science particularly since student performance is poor in these subject areas. Sukkur IBA has recommended to the World Bank that the data collected be made public so that other academics and research organizations are able to study it further and design meaningful interventions on the basis of it. Data as it is reported at present (i.e. in the form of averages across subjects and districts) is of limited use.

KHYBER PAKHTUNKHWA

Developing and disseminating results

Given that large-scale assessments are relatively new in Khyber Pakhtunkhwa (KP), it is in the process of developing its capacity and a strategy for disseminating assessment results to drive its quality reform agenda. As such, it is not surprising if early efforts to conduct large-scale assessment are not accompanied by an effective communication strategy. Results from the 2015 large-scale assessment of grade 5 students were not disseminated to all stakeholders. For the 2016 cycle, the Peshawar Board of Intermediate and Secondary Education (BISE) is in the process of compiling the results and will share them with the Directorate of Curriculum and Teacher Education (DCTE) for analysis. Plans are afoot to feed the results of the analysis, as soon as they are available, into the planning process of teachers' professional development. Whether these results will be used depends on the level of synergy between these assessments efforts, which are driven primarily by policymakers in the Elementary and Secondary Education Department, and other policymakers/stakeholders in the system, especially since there is currently no accompanying legislation mandating it.

The capacity for assessment data analysis keeps fluctuating at Provincial Education Assessment Center (PEAC) in KP. At present, the trained staff has been transferred out of PEAC to other institutions and untrained staff has been posted in its place. Given the drop in its capacity, due to loss of trained human resources, PEAC is currently relying upon external consultants for analysis of assessment data. It is also relying on external technical assistance to develop other materials such as calendars displaying assessment results for teachers. Results were shared by PEAC at several levels. At the school level, results were shared in the form of calendars and posters which contained challenging SLOs, proposed activities that teachers could carry out with students, and some broader recommendations for improvement in weak areas. At the district level, seminars were held with district managers. PITE officials and other teachers were also informed of results in targeted meetings. The main report on the sample-based assessment results was shared with the Secretary and other stakeholders.

Using results

Plans are afoot to turn the low-stakes large-scale assessment into examinations from the next cycle onwards (i.e. in 2017). Once in place, the examination results will be used to promote students to the next grade on the basis of their performance. At present, the assessment data is used for informational purposes only.

Although PEAC has done a great deal to ensure that results of the sample-based assessment are disseminated, their results are for the most part not used.[13] In fact, PEAC has struggled with determining strategies for ensuring greater use. In lieu of this, PEAC is planning to use the 2016 sample-based assessment results to inform the prioritization and planning of teachers' professional development. PEAC is part of the DCTE and the latter has the primary responsibility to manage professional development within the province. The close proximity will make it easier for PEAC to provide timely feedback to DCTE for use in yearly planning of teachers' professional development.

SECONDARY AND HIGHER SECONDARY LEVEL EXAMINATIONS

BOARDS OF INTERMEDIATE AND SECONDARY EDUCATION

Developing and disseminating results

Activity at the secondary level is limited to tabulation and reporting of raw exam results for individual candidates only, which are made available online and through a printed gazette. There is no further analysis of results and production of reports.

In instances where data is computerized (e.g. in Lahore, Peshawar, and Karachi), it is usually used to verify the handwritten score (i.e. that all questions have been marked and computed correctly) rather than for data analysis. The boards in Lahore and Karachi have done some descriptive analyses with the data available (i.e. by gender, academic groups, or over time). However, beyond presentations to government officials, it is not clear how this data is used. It is also not publicly available. The only example of analysis of the examination data is a recent report of the Quetta BISE data developed with donor and civil society support which looked at trends over the years, by gender, and type of school.[14] This was meant to be a demonstration of what can be done with the results. However, it is not clear whether this exercise would be replicated by the Quetta BISE in the future.

The limited reporting of exam results is not surprising as there is no clear mandate for data analysis or research in the BISE. Still, the lack of these activities are due to the lack of departments in some cases in smaller boards, unfilled positions, lack of adequate personnel with research backgrounds, and most importantly, lack of funds to conduct research activities. The research sections that do exist are also predominantly involved in non-research activities such as issuing scholarships and arranging debates or competitions for students.

Using results

By and large, the BISE examination results are not used to inform the learning process or policy. Results are primarily used for student promotion into the next grade or for certification. However, universities do not appear to trust the quality of exams and give their own entrance exams. The National Testing Service also administers a university entrance exam. Most recently, the Higher Education Commission has set up an Educational Testing Council to administer admissions tests- both public and private universities will be required to use these, with individual universities responsible for determining the weight given to the test in student admissions.[15] Results are also used for teacher evaluation at the secondary level and can be grounds for determining dismissal, promotion, or even incentives in Punjab.[16]

AGA KHAN UNIVERSITY-EXAMINATION BOARD

Developing results

Aga Khan University-Examination Board (AKU-EB) results include: (a) an overall distribution of grades in the school; (b) a distribution of grades across the sciences and humanities; (c) a subject-wise distribution of grades; (d) a subject-wise distribution of scoring ranges with intervals of five marks; and (e) item-level scores for multiple choice questions, constructed response questions, and extended response questions across the cognitive categories of knowledge, understanding, and application. Additionally, the report shows the three-year trend of total grade distribution and lists the names of schools in which at least two-thirds of candidates secured 'A' grades or above in compulsory subjects.

Disseminating results

After exams are conducted, results are announced online within six weeks. After a further ten days, transcripts are given to students and certificates are issued afterwards so students can apply to undergraduate programs. As AKU-EB is recognized as an independent board and as a member of the Inter Board Committee of Chairmen by the 2002 Ordinance, AKU-EB students do not need equivalency certificates while applying to such programs.

Individualized School Performance Reports (SPRs) are generated and sent to each school affiliated with the board. These reports compare the school's performance in both parts of the Secondary School Certificate examination with national results, and with performance from previous years in multiple ways (refer to Figures 6.1 and 6.2). SPRs are shared with content experts to compare performance across regions, and the same report is shared with item and curriculum developers.

Using results

AKU-EB results are only used by its affiliated schools and by AKU-EB itself. Some schools have used the SPRs as appraisals for their teachers.[17] If the report shows that a substantial proportion of students performed poorly in a subject, results are flagged with an asterisk, indicating that the average is low because of the poor performance of the entire class. These results shape the focus of AKU-EB teacher training workshops for the school as well.

Figure 6.1 Comparison of school results with national results as reported in the School Performance Report (SPR)
Source: Adapted from AKU-EB affiliated school's SPR (2015)

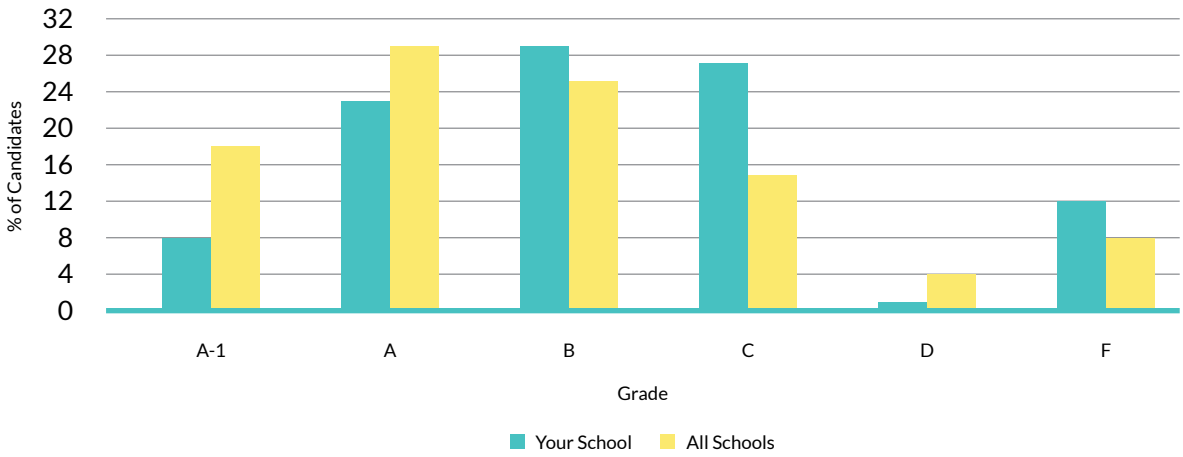
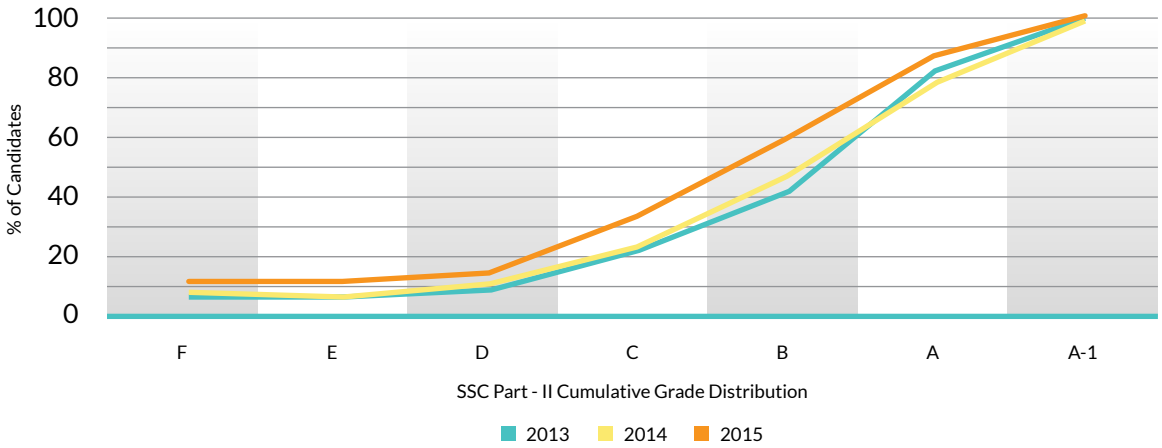


Figure 6.2 Combined grade distribution over time as reported in the School Performance Report (SPR)
Source: Adapted from AKU-EB affiliated school's SPR (2015)



CONCLUSION

The priority that is given to the analysis, communication, and use of results varies across assessments in Pakistan. Once again there is a bifurcation, with the primary and elementary level assessments and AKU-EB placing greater emphasis on result production and communication as opposed to the BISE (see Table 6.5). There continues, however, to be room for improvement across the different levels of assessment with capacity often lacking in assessment agencies in this critical area.

The foregoing description demonstrates that teaching and learning practices in classrooms, textbook development processes, and on-going professional development of teachers remain largely uninformed by assessment results. There is limited premium on the use of assessment data in determining which aspects of the education system need to be improved and how. As such, the practices of designing and implementing assessments show positive growth and improvement over time. However, there is no robust policy enabling the use of assessment results to drive improvement in the system. In some cases, it also seems that the (perceived) lack of credibility of the assessment hampers use of results.

Challenges exist at both ends, that is, the production and communication of results, as well as the use of results. Results need to be produced in a precise and simple manner and communicated to stakeholders in time. At present, assessment results reports take a long time to be produced and disseminated due to lack of capacity within assessment agencies to undertake these activities efficiently. Even when ready, the reports are often inaccessible. With regard to the use of assessments results, planners of teachers' professional development, district managers, and school administrators need to be sensitized, and perhaps even trained on how to make use of the data and results to inform their work, thereby, improving quality in the sector.

Policy makers should make use of data and information on student learning generated from large-scale assessments with great care and caution. There is a tendency to use the results as a proxy for teacher performance, which, as discussed in detail above, is not a justifiable policy. Teachers need to be held accountable; however, assessment results are only one source of information amongst many that should be used to measure their performance. Assessments are a worthwhile exercise if results are produced, communicated and used to engage in dialogue and take action to improve various aspects of the education system and ultimately student learning.

Table 6.5 Production and dissemination of assessment results in Pakistan

	PEC	SAT	KP G5[18]	BISE	AKU-EB
Production of student results	✓	✓	n/a	✓	✓
Production of provincial and/or district level results	✓	✓	n/a	×	n/a
Use of multiple indexes in results	✓	✓	n/a	×	✓
Use of proficiency levels in results	×	×	n/a	×	×

[1] Cresswell, Schwantner & Waters (2015); Kellaghan, Greaney & Murray (2009); Tobin, Lietz, Nugroho, Vivekanandan, & Nyamkhuu (2015)
[2] Cresswell, Schwantner & Waters (2015); Kellaghan, Greaney, & Murray (2009)
[3] Best, Knight, Lietz, Lockwood, Nugroho & Tobin (2013)
[4] Kellaghan, Greaney & Murray (2009); Tobin et al. (2015)
[5] Kellaghan, Greaney & Murray (2009)
[6] The DSD is responsible for the in-service training of all primary school teachers in the public sector in the Punjab province.
[7] Government of Punjab Notification (2011, April 29). Performance evaluation was calculated on the basis of exam results (70%), test participation rate (15%) and gain in enrollment (15%).
[8] Malik (2013)
[9] RSU & Sukkur IBA (2015)
[10] Sukkur IBA staff interview (2 December 2015)
[11] RSU staff interview (30 November 2015)
[12] Sukkur IBA staff interview (2 December 2015)
[13] PEAC KP staff interview (26 February 2016)
[14] Alif Ailaan and SCSPEB (2015)
[15] "HEC sets up Education Testing" (2016)
[16] Government of Punjab Notification (2011, April 29)
[17] AKU-EB staff interview (4 December 2015)
[18] The 2016 assessment cycle was in process during the production of this report therefore there is insufficient information available to report on these indicators

07

CONCLUSION

THE WAY FORWARD

Formal government sponsored large-scale and sample-based assessments of student learning are a relatively recent phenomenon in Pakistan. Indeed, it would not be entirely inaccurate to say that nearly three decades ago, the term 'assessment' was not even used in Pakistan with reference to student learning. Schools had a system of quarterly, six-monthly, and annual examinations. In the 1990s, the discourse and practice of standards and standardized assessment began to stream into Pakistan in the wake of global education reforms movements. By the beginning of the 21st century, regular student assessments had become the mainstay of education policy. Governments, development partners, civil society, academia, nearly everyone wanted to know and have various aggregated and disaggregated statistics about student learning achievements. Assessments have been used for comparison of districts, provinces, and between public and private schools, to study the impact of particular interventions in education, and to reward or penalize teachers. There are other ways in which results of assessments can be, but have not been, used such as to inform and drive improvements in the professional development of teachers. This report has looked at the history, enabling environment, design, and implementation of these assessments as well as the dissemination and use of results.

The chapters in this report followed a common pattern of describing best practices for each step in the assessment cycle followed by a review of the actual practices in various provinces of Pakistan. The report finds that the enabling conditions for high quality large-scale assessments vary across the provinces. Enabling conditions comprise a mix of legislation, institutions, and favorable policy stances. The ideal scenario for design, conduct, and use of assessments would be having an assessment agency that is fully supported by appropriate legislation and robust policy on the use of assessment to drive improvements in the education system. However, not all of these conditions are fulfilled at the same time in any province. In Punjab, legislation and an assessment agency exist, but a policy on the assessments that the province conducts regularly does not exist. Sindh has neither legislation nor an institutional framework to lend support to its assessment activities. Khyber Pakhtunkhwa is struggling to use its current institutional framework to design, as well as conduct assessments. While it has a well-defined policy regarding the use of assessment data, there is no legislation.

There are also substantive differences between the ways in which assessments are conceptualized at the primary and elementary levels, and the secondary and higher secondary levels. While the former subscribe willingly to more modern approaches to assessment, the latter have been grounded in the traditional approach to the development and conduct of examinations. Assessments drive teaching and learning practices in classrooms. Their reform can have a feedback effect on teaching and learning. Therefore, to derive maximum benefit from large-scale assessments, the provinces will need to enforce uniform standards for assessment at all levels of schooling. Sooner or later, the secondary and higher secondary examinations will need to follow the standards and established best practices for design and implementation of assessments. Given below are some specific recommendations based on the discussions in various chapters of this report.

RECOMMENDATIONS

ENABLING ENVIRONMENT

Develop national and provincial policies for assessment along with legislative cover

Currently, there are a variety of assessments in operation in the provinces and at the national level. Each assessment has its own purpose and provides different kinds of information on student learning outcomes. For example, National Education Assessment System has reemerged; however, it is not clear how the findings of the national sample-based assessment will complement the efforts of large-scale assessments being conducted at the provincial level. It is imperative that the provinces and the federal government deliberate on what mix of assessments is needed to fulfill government objectives and information needs. Coordinated assessment policies will help provide greater clarity on how these goals are being met, ensure more efficient use of resources, and prevent assessment overload for students and teachers.

This process of policy development must go hand in hand with legislation. In certain cases where policies exist, legislative cover does not. Both are needed and are important to ensure the stability and continuity of assessments.

Restructure assessment agencies and improve capacity of human resources

Although the large-scale assessment agencies at the primary and elementary levels have focused on developing the capacity of their human resources, it still remains scarce. Human resource development within assessment agencies should be the top priority of provincial governments. A major challenge is lack of technical staff to lead test construction and data analysis. A carefully thought out strategy involving both higher education institutions and assessment agencies is needed. Higher education institutions must be able to produce the much needed number of assessment professionals. They need to introduce courses in educational measurement and evaluation which should be picked up by assessment agencies. Meanwhile, the assessment agencies, including the BISE, should reconsider their structure to include technical departments and technical positions where needed. Preference should be given to assessment experts rather than bureaucrats. They should then hire professionals as soon as they become available and train them further through a professional development strategy. Finally, there is also a need for more rigorous selection and training of temporary staff, such as item writers and scoring staff.

Introduce courses on assessment and measurement

As mentioned above, technical professionals with appropriate preparation to design, score, and analyze assessments are missing in the market altogether. Opportunities to study educational measurement and evaluation at the university level are limited and are often not relevant to the needs of modern assessments. Assessment agencies may need to determine and communicate their needs to the universities so that relevant high quality programs can be designed. The Higher Education Commission can play a coordinating role in determining the number of assessment professionals needed and call upon the universities to respond to this need.

Align assessments with curriculum, professional development and other quality aspects

Alignment is key to assessment reforms as well as education reforms in general. Assessment must align with the curriculum. The alignment of instruction with curriculum will automatically follow. Teaching to the test should not be seen as a big issue. It is actually good to teach to high quality tests. However, alignment of teaching with low quality assessments can drive down learning. Furthermore, teacher preparation and professional development will be more responsive to the learning needs of students if they are informed by data generated by regular assessments. Emphasis on timely communication of assessment results to professional development agencies and the use of those results by the latter to prioritize training is a central message of this report.

Reduce the number of examination boards or create an apex board in each province

Better quality of examinations within each province will be ensured if the papers are set by a single board of examination instead of multiple boards. In Pakistan, the number of boards has multiplied over time. This increase in the number of boards has been driven by political rather than technical considerations. In other countries, there is often one board per state (such as in India) or for the entire country (such as in Singapore, Malaysia, and New Zealand). Only where there is competition between the private test publishers (such as in the United Kingdom) do multiple boards exist. Therefore, reducing the number of boards, as also suggested in the National Education Policy 2009, may help make better use of scarce human resources available in the country. Ideally, there should be one apex board in each province that is equipped to design examinations according to accepted standards and the remaining boards should only administer the exams designed by the apex board.

Introduce school-based assessment

Education policies in Pakistan have time and again noted the need for school-based assessment to complement external examinations. Increasingly, assessment systems around the world have moved towards including school or classroom-based assessments into the mix of assessments they conduct. School-based assessments by their very nature are more localized which, particularly if they are formative, have the advantage of allowing teachers to get the feedback they need in real-time, and inform their teaching practice. They also allow for room to assess a wider range of cognitive skills and other non-cognitive skills that cannot be tested through summative assessment or examinations. However, there are significant challenges in implementing school-based assessment policies, particularly in the developing context. They require heavy investments in teacher education and professional development, something which may not be feasible in the short term. Therefore, there is a need for a nuanced policy in this regard- develop teacher capacity to design and interpret assessment results in the long term with the ultimate goal of handing a portion of the assessment responsibility over to the school and teacher.

ASSESSMENT PRACTICES

Institute standards in assessment design

In order to ensure valid, reliable, and fair assessments, it is critical that assessment design follows internationally accepted standards for design. This entails following all steps in the cycle: developing a test specifications document, training item writers to understand the characteristics of good items, writing and reviewing items multiple times to ensure adherence to test specifications and psychometric criteria, piloting test items on a representative sample of students, conducting psychometric analysis on the pilot data to select final items for the assessment, and putting together technical documentation on the steps involved in designing the assessment. This is highly technical work requiring staff with requisite capacity. As noted earlier, the recently established assessment agencies, mostly at the primary and elementary level, are more amenable to adhering to new ways of designing assessments. However, it is counterproductive to change practices at these levels while leaving them untouched at the secondary and higher secondary levels. A concerted reform effort is needed at all levels.

Follow best practices in implementation of assessments

As with design, it is important that the implementation of assessments also follow internationally accepted standards. In the administration of assessments, there are two areas that require particular attention: lack of preparation on the part of invigilators, and use of practicing teachers to administer assessments. For the former, training is required so that all administrative staff is adequately prepared. With reference to the use of practicing teachers, this can be especially problematic in the case of high-stakes assessments. Governments should explore the possibility of making use of retired teachers in the administration of such assessments. With regard to the scoring process, much work is needed. In order to improve scoring accuracy, prevent unfair means, and ensure greater efficiency, Optical Mark Recognition (OMR) software for scoring of multiple choice questions needs to be made use of. For marking of constructed response questions, once again, scoring staff need to be trained and provided detailed rubrics. Syndicate scoring, where possible, should take place and multiple levels of review of the scoring process, for greater quality assurance, need to be instituted.

Improve transparency, leakage and cheating accountability

The use of unfair means, especially in high-stakes assessments, abound. Accountability mechanisms pertaining to personnel deployment and performance need to be put in place. Experience suggests that development of multiple versions of papers and use of OMR software to score papers can go a long way in reducing the use of unfair means and should be considered in all provinces.

Effectively disseminate assessment results to ensure their use

It is clear that data is either not produced or when it is produced is not utilized to inform decisions, policy, or teaching practice. Assessment agencies need to have the capacity to analyze data, develop results, and write up reports. Most importantly, they need to have an effective communication strategy, in which results are written up differently for different stakeholders. For example, schools, teachers, and professional development institutions would be more interested in student performance by cognitive domains or learning outcomes, whereas policymakers or the media would be more interested in brief summary statistics of performance by different regions or groups. Assessment results must be precise, simple, and responsive to the needs of ultimate users and be delivered in a timely manner.

Deliberate on alternative means for determining teacher accountability

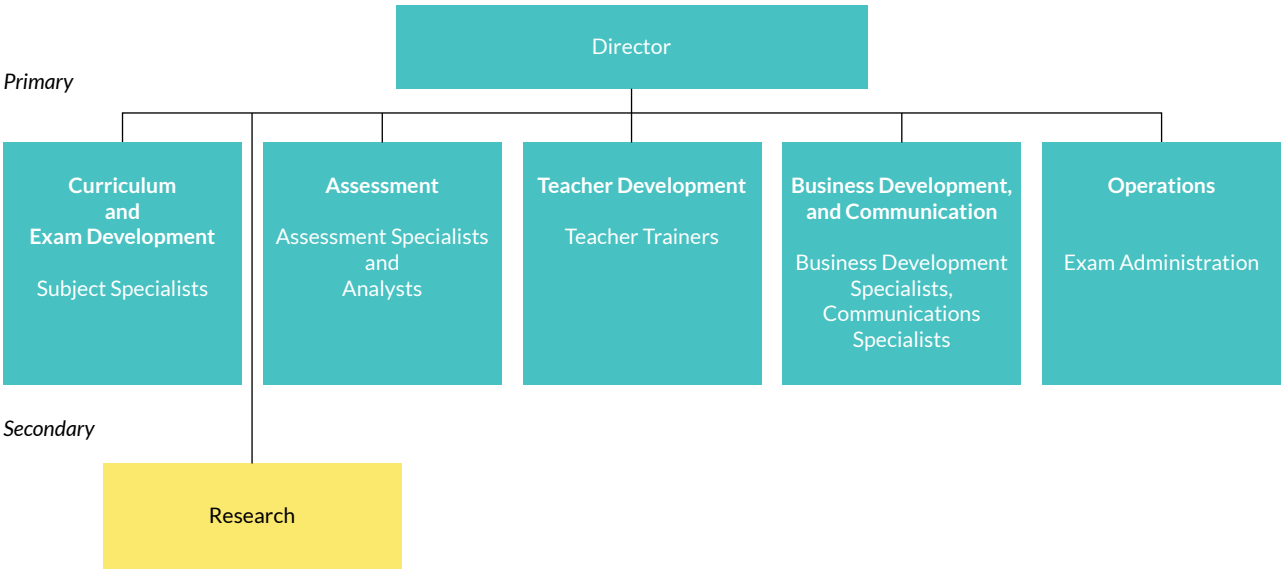
In previous chapters, the disadvantages of using assessment results for teacher accountability have been discussed. Student performance can be attributed to a variety of factors in addition to the teacher. While teacher accountability is important, it would be best if policymakers deliberated on other means for determining teacher performance. Teacher accountability can be based on a wide array of indicators rather than solely on student performance on an assessment.

To sum up, assessments have a key role in driving quality in the education system. As such, much is to be gained from spending considerable time and effort in improving assessment systems. Among other steps, this requires the elaboration of an assessment policy, the provision of legislative cover, and institutional strengthening. High quality testing is very likely to have a positive effect on the quality of teaching and learning. However, there is room for discussion on the grade and age level at which such tests serve a useful purpose and the frequency with which they ought to be administered. Certainly, the use of student performance in such tests as a key accountability mechanism is a subject of considerable debate and concern, globally. That having been said, there is little doubt as to the central role of assessment systems in the context of quality in any education system. Given that, the governments should attend to the task of reforming it with the urgency it deserves.

APPENDIX

Figure A: BISE Bahawalpur organogram

Figure B: AKU-EB organogram



REFERENCES

5th, middle class exams: invigilators' remuneration increased. (2016, February 8). Pakistan Today. Retrieved from http://www.pakistantoday.com.pk/2016/02/08/national/5th-middle-class-exams-invigilators- remuneration-increased/

8th class maths paper leaked. (2014, February 14). Dawn. Retrieved from http://www.dawn.com/news/1086892/8th-class-maths-paper-leaked

9th Class Annual Exam results announced. (2015, August 22). The Nation. Retrieved from http://nation.com.pk/national/22-Aug-2015/9th-class-annual-exam-results-announced

Across the metropolis: Despite secondary board's shortcomings, exams begin today. (2016, March 28). The Express Tribune. Retrieved from http://tribune.com.pk/story/1073990/across-the-metropolis-despite-secondary-boards-shortcomings-exams-begin-today/

AKU-EB (2012). Secondary School Certification Examination Syllabus, Biology. Retrieved from http://examinationboard.aku.edu/forstudents/Documents/Biology%20_Classes%20IX-X_NC%202006_%20Latest%20Revision%20June%202012.pdf

Ali, R. (2012, May 9). Chalk and talk is out, says private board slowly working on unlearning bad habits. The Express Tribune. Retrieved from http://tribune.com.pk/story/376079/chalk-and-talk-is-out-says-private-board-slowly-working-on-unlearning-bad-habits/

Alif Ailaan & SCSPEB. (2015). Pass/Fail? Matriculation examination results in Balochistan and what they mean for the future. Islamabad: Alif Ailaan.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). The standards for educational and psychological testing. Washington, DC: AERA.

Anderson, P., & Morgan, G. (2008).Developing tests and questionnaires for a national assessment of educational achievement. In V. Greaney & T. Kellaghan (Ed.), National Assessments of Educational Achievement(Vol. 2). Washington, DC: World Bank.

Basu, B. D. (1867). History of education in India under the rule of the East India Company: Modern Review Office.

Best, M., Knight, P., Lietz, P., Lockwood, C., Nugroho, D., & Tobin, M. (2013). The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries. Final report. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Board of intermediate and Secondary Education Sahiwal. (n.d.). Retrieved from http://www.bisesahiwal.edu.pk/who-we-are.php

Carlson, J. E., & Davier, M. (2013). Item Response Theory. ETS Research Report Series, 2013(2), i-69.

Christie, T., & Afzaal, M. (2005). Achievement in Pakistan: An empirical investigation of a widespread assumption, paper presented at an IAEA international conference. Abuja, Nigeria.

Clarke, M. (2012). What matters most for student assessment systems: A framework paper. Washington, DC: World Bank.

CM orders probe into grade 5 English leak. (2014, February 11). The News International. Retrieved from http://www.thenews.com.pk/Todays-News-5- 231750-CM-orders-probe-into-grade-5-English-paper-leak

Cresswell, J., Schwantner, U., & Waters, C. (2015). A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data, PISA. Washington, DC: World Bank.

de Castro, M. H. G. (2012). Developing the enabling context for student assessment in Brazil. Washington, DC: World Bank.

DFID. (2014). PESP II Annual Review.

Education boards in Zhob, Turbat planned . (2011, October 30). Dawn. Retrieved from http://www.dawn.com/news/670178/education-boards-in-zhob-turbat-planned

Ferrer, J. G. (2006). Educational assessment systems in Latin America: Current practice and future challenges. Washington, DC: Preal.

Government of N.W.F.P. (1995). The Calendar of the Boards of Intermediate and Secondary Education, N.W.F.P. Peshawar.

Government of Pakistan, Ministry of Education. (1972). The Education Policy 1972-1980. Islamabad: Government of Pakistan.

Government of Pakistan, Ministry of Education (1998). National Education Policy 1998-2010. Islamabad: Government of Pakistan.

Government of Punjab Notification. (2011, April 29).

Government of Sindh, Education and Literacy Department. Reform Support Unit (RSU). (2011). Sindh Provincial Standardized Assessment Test (SATs) Framework.

Government of Sindh, Education and Literacy Department. Reform Support Unit (RSU), & Sukkur IBA. (2012). Statistical Analysis of SAT-1 (2012) results.

Government of Sindh, Education and Literacy Department. Reform Support Unit (RSU). (2013). Sindh Education Sector Plan 2013-2016.

Greaney, V., & Kellaghan, T. (2008). Assessing national achievement levels in education. In V. Greaney & T. Kellaghan (Ed.), National Assessments of Educational Achievement (Vol. 1). Washington, D.C: World Bank.

Greaney, V., & Kellaghan, T. (2012). Implementing a national assessment of educational achievement. In V. Greaney & T. Kellaghan (Ed.), National Assessments of Educational Achievement (Vol.3). Washington, D.C: World Bank.

HEC sets up Education Testing Council. (2016, April 16). Dawn. Retrieved from http://www.dawn.com/news/1252373

Institute of Social and Policy Sciences (I-SAPS). (2015). Public Financing of Education in Punjab: Provincial and District Level Analysis. Islamabad: I-SAPS.

Kanjee, A., & Acana, S. (2013). Developing the Enabling Context for Student Assessment in Uganda. (SABER – Student Assessment Working Paper No. 8). World Bank: Washington, DC.

Kellaghan, T., Greaney, V., & Murray, S. (2009). Using the results of a national assessment of educational achievement.In V. Greaney & T. Kellaghan (Ed.), National Assessments of Educational Achievement (Vol.5). Washington, D.C: World Bank.

Malik, M. (2013, August 5). Teachers and 'conundrum' of poor results. Dawn. Retrieved from http://www.dawn.com/news/1034154

Ministry of Education. (1953). Report of the Secondary Education Commission: Mudaliar Commission Report. Ministry of Education, Government of India.

Ministry of Education. (1959). Report of the Commission on National Education. Karachi: Government of Pakistan Press.

Moses, M., & Nanna, M. (2007). The Testing Culture and the Persistence of High Stakes Testing Reforms. Education and Culture,23(1), 55-72. Retrieved from http://www.jstor.org/stable/42922602

Mujeeb, A. M., Pardeshi, M. L., & Ghongane, B. B. (2010). Comparative assessment of multiple choice questions versus short essay questions in pharmacology examinations. Indian journal of medical sciences, 64(3), 118.

Mullis, I. V., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004).TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. TIMSS & PIRLS International Study Center.

New BISE on the cards. (2015, September 11). Dawn. Retrieved from http://www.dawn.com/news/1206147

Ordinance No. CXIV of 2002. An Ordinance to establish the Aga Khan University Examination Board. (2002). Retrieved from http://nasirlawsite.com/laws/akuebo.html

Oxford and Cambridge started school examinations in 1858. Riding, R. J., & Butterfield, S. (1990). Assessment and examination in the secondary school: a practical guide for teachers and trainers: Taylor & Francis.

PPIU, Education Department, Government of Balochistan (2013). BalochistanEducation Sector Plan 2013-2017.

Punjab Examination Commission (PEC). (2015). Report on Grades 5 and 8 Exams.

Report of the Commission on National Education. (1959). Karachi: Government of Pakistan Press.

Riding, R. J., & Butterfield, S. (1990). Assessment and examination in the secondary school: a practical guide for teachers and trainers: Taylor & Francis.

Sacks, P. (1997). Standardized Testing: Meritocracy's Crooked Yardstick. Change,29(2), 24-31. Retrieved from http://www.jstor.org/stable/40165509

Sadler, S. M. E. (1919). Report of the Calcutta University Commission (1917-1919). Calcutta: Government of India.

Saleem, M. (2007). Remarking of education in Pakistan: A critique of Aga Khan University education board. The Dialogue, 2(1), 27-64.

Shah, D., & Afzaal, M. (2004, June). The examination Board as Educational Change Agent: The Influence of question choice on selective study, paper presented at 30th annual IAEA Conference. Philadelphia, USA.

Sharma, O. P. (1991). Administration of education boards in India: APH Publishing.

Society for the Advancement of Education (SAHE).(2011). Improving education through large-scale testing? A study on primary and elementary level exams in Punjab. Lahore: SAHE.

Timrazi, S. (2008). Consultancy Report, Institutional Strengthening of the National Education Assessment System Network (NEAS, PEACEs and AEACs).

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region. Melbourne: ACER.

United States Agency for International Development (USAID). (2008). Evaluation for the Aga Khan University Examination Board (AKU-EB). Retrieved from http://pdf.usaid.gov/pdf_docs/Pdacl727.pdf

Verger, A., Altinyelken, H., & De Koning, M . (Eds.). (2013). Global managerial education reforms and teachers: emerging policies, controversies and issues in developing countries.. Education International Research Institute IS Academic Program.

World Bank. (2004). Program Document for Punjab Education Sector Adjustment Credit.

World Bank. (2006). Program Document for Third Punjab Education Development Policy Credit.

World Bank. (2007). Information Completion Report for Fourth Punjab Education Development Policy Credit.

World Bank. (2012a). Implementation Completion and Results Report for a Sindh Education Sector Project. Retrieved from http://documents.worldbank.org/curated/en/2012/12/17155963/pakistan-sindh-education-sector-project

World Bank. (2012b.) Implementation Completion and Results Report For Punjab Education Sector Project (PESP).

World Bank. (2012c). Project Appraisal Document For a Second Punjab Education Sector Project.

World Bank. (2012d). SABER Province Report Sindh, Pakistan: Student assessment.

World Bank. (2013). Project Appraisal Document For a Second Sindh Education Sector Project.

Young, J. W., & Kobrin, J. L. (2001). Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2001-6-differential-validity-prediction-college-admission-testing-review.pdf

Zia, A. (2016, January 17). Grade-5 exams: Educationists demand Pashto language paper for 5th graders. The Express Tribune. Retrieved from http://tribune.com.pk/story/1028740/grade-5-exams-educationists-demand-pashto-language-paper-for-5th-graders/

Society for the Advancement of Education

Campaign for Quality Education

OPEN SOCIETY FOUNDATIONS