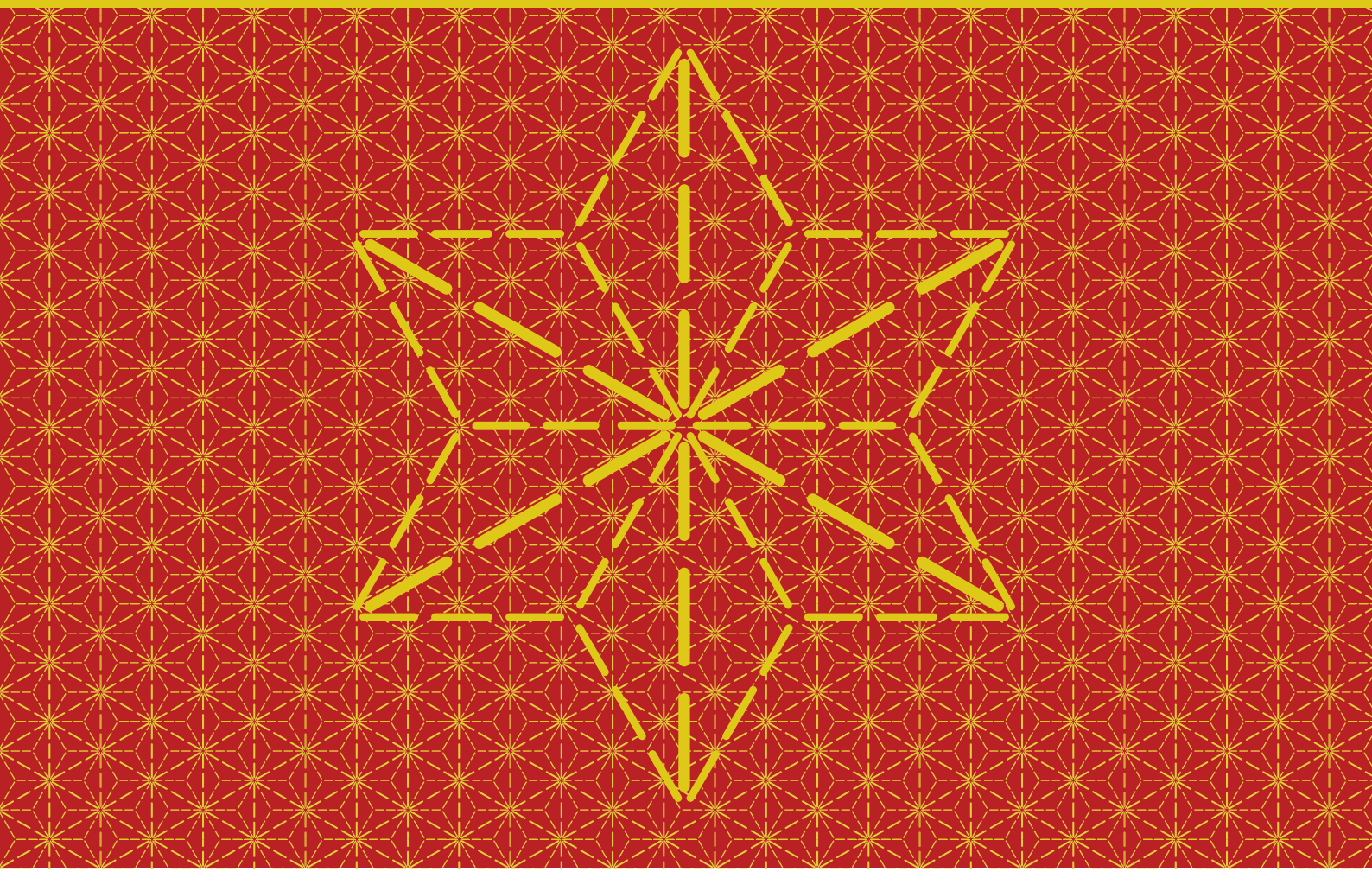
semester 8

Test Development and Evaluation

WINDOWS ON PRACTICE GUIDE

B.Ed. (Hons.) Elementary

2012



Higher Education Commission

This product has been made possible by the support of the American People through the United States Agency for International Development (USAID). The contents of this report are the sole responsibility of the authors, and do not necessarily reflect the views of USAID or the United States Government.

Technical Support: Education Development Center (EDC); Teachers College, Columbia University

Higher Education Commission

Foreword

Teacher education in Pakistan is leaping into the future. This updated Scheme of Studies is the latest milestone in a journey that began in earnest in 2006 with the development of a National Curriculum, which was later augmented by the 2008 National Professional Standards for Teachers in Pakistan and the 2010 Curriculum of Education Scheme of Studies. With these foundations in place, the Higher Education Commission (HEC) and the USAID Teacher Education Project engaged faculty across the nation to develop detailed syllabi and course guides for the four-year B.Ed. (Hons) Elementary and the two-year Associate Degree in Education (ADE).

The syllabi and course guides have been reviewed by the National Curriculum Review Committee (NCRC) and the syllabi are approved as the updated Scheme of Studies for the ADE and B.Ed. (Hons) Elementary programmes.

As an educator, I am especially inspired by the creativity and engagement of this updated Scheme of Studies. It offers the potential for a seismic change in how we educate our teachers and ultimately our country's youngsters. Colleges and universities that use programmes like these provide their students with the universally valuable tools of critical thinking, hands-on learning, and collaborative study.

I am grateful to all who have contributed to this exciting process; in particular the faculty and staff from universities, colleges, and provincial institutions who gave freely of their time and expertise for the purpose of preparing teachers with the knowledge, skills, and dispositions required for nurturing students in elementary grades. Their contributions to improving the quality of basic education in Pakistan are incalculable. I would also like to thank the distinguished NCRC members, who helped further enrich the curricula by their recommendations. The generous support received from the United States Agency for International Development (USAID) enabled HEC to draw on technical assistance and subject-matter expertise of the scholars at Education Development Center, Inc., and Teachers College, Columbia University. Together, this partnership has produced a vitally important resource for Pakistan.

PROF. DR SOHAIL NAQVI
Executive Director
Higher Education Commission
Islamabad

How the Windows on Practice guide was developed

As part of nationwide reforms to improve the quality of teacher education, the Higher Education Commission (HEC), with technical assistance from the USAID Teacher Education Project, engaged faculty across the nation to develop courses for the new four-year B.Ed. (Hons) Elementary programme.

The process of designing the syllabus for each course in years 3–4 of the programme began with curriculum design workshops. Faculty who will teach the courses were identified by university deans and directors and then invited to attend the workshops. The first workshop included national and international subject matter experts who led a seminar focused on a review and update of subject (content) knowledge. The remainder of this workshop was spent reviewing the HEC Scheme of Studies, organizing course content across the semester, developing detailed unit descriptions, and preparing the course syllabi. Although the course syllabi are designed primarily for Student Teachers taking the course, they are a useful resource for teacher educators, too.

Following the initial workshop, participants developed teaching notes, including ideas for teaching units of study and related resources. Faculty worked individually or in groups, focusing on their own preparations to teach, while bearing in mind that their end product must also be useful to those who will teach the course in the future. A series of workshops occurred over the year in order to allow faculty to have time for their work, engage in peer review, and receive critical feedback from national and/or international consultants. In designing both the syllabi and the teaching notes, faculty and subject matter experts were guided by the National Professional Standards for Teachers in Pakistan 2009.

All of the syllabi developed by faculty are included in this document along with a listing of topical teaching notes. Additional references and resources appear at the end of the document. These should provide a rich resource for faculty who will teach the course in the future. Sample syllabi with accompanying teaching notes are also included in order to provide new faculty with a model for developing curriculum and planning to teach. This Windows on Practice guide is not intended to provide a complete curriculum with a standard syllabus and fully developed units of study, rather it aims to suggest ideas and resources for faculty to use in their own planning. Hence, readers will find sample units and materials that reflect the perspective of faculty designers rather than prescriptions for practice.

We are respectful of intellectual property rights and have not included any suggested materials that are copyright protected and for which we have not secured explicit permission to use. Therefore, all materials included may be used in classrooms for educational purposes. Materials in this document are not intended for commercial use, however. They may not be used in other publications without securing permission for their use.

Initial drafts were reviewed by the National Curriculum Review Committee (NCRC), and suggestions were incorporated into final drafts, which were then submitted to the NCRC for approval.

Faculty involved in course design: Dr Bashir Hussain, Bahauddin Zakariya University, Multan; Ishrat Siddiqa Lodhi, Fatima Jinnah Women University, Rawalpindi; Dr Khalid Saleem, University of Education, Lahore; Dr Muhammad Sarwar, University of Sargodha; Naila Siddiqua, University of Karachi; Dr Naveed Sultana, Allama Iqbal Open University, Islamabad; Raqeeb Imtiaz, University of Gujrat; Dr Rizwan Akram Rana, University of Punjab, Lahore; Sadruddin Qutoshi, Karakorum International University, Gilgit; Dr Saifullah, University of Gujrat, Lahore; Saima Mushtaq, Fatima Jinnah Women University, Rawalpindi; Saira Soomro, University of Sindh; Samreen Jalal, University of Education Lahore; Dr Shafqat Hussain, University of Sargodha; Zahid Majeed, and Allama Iqbal Open University, Islamabad.

National subject experts leading the seminar: Dr Thomas Christie, Aga Khan University Examination Board (AKU-EB), Karachi, and Isbah Mustafa, AKU-EB, Karachi.

NCRC review dates: 24–25 April 2013

NCRC Reviewers: Dr Saeed Khan, University of Haripur, Khyber Pakhtunkhwa; Dr Sabiha Hammid, Islamia University, Bahawalpur; and Dr Muhammad Imran, Arid Agriculture University, Rawalpindi.

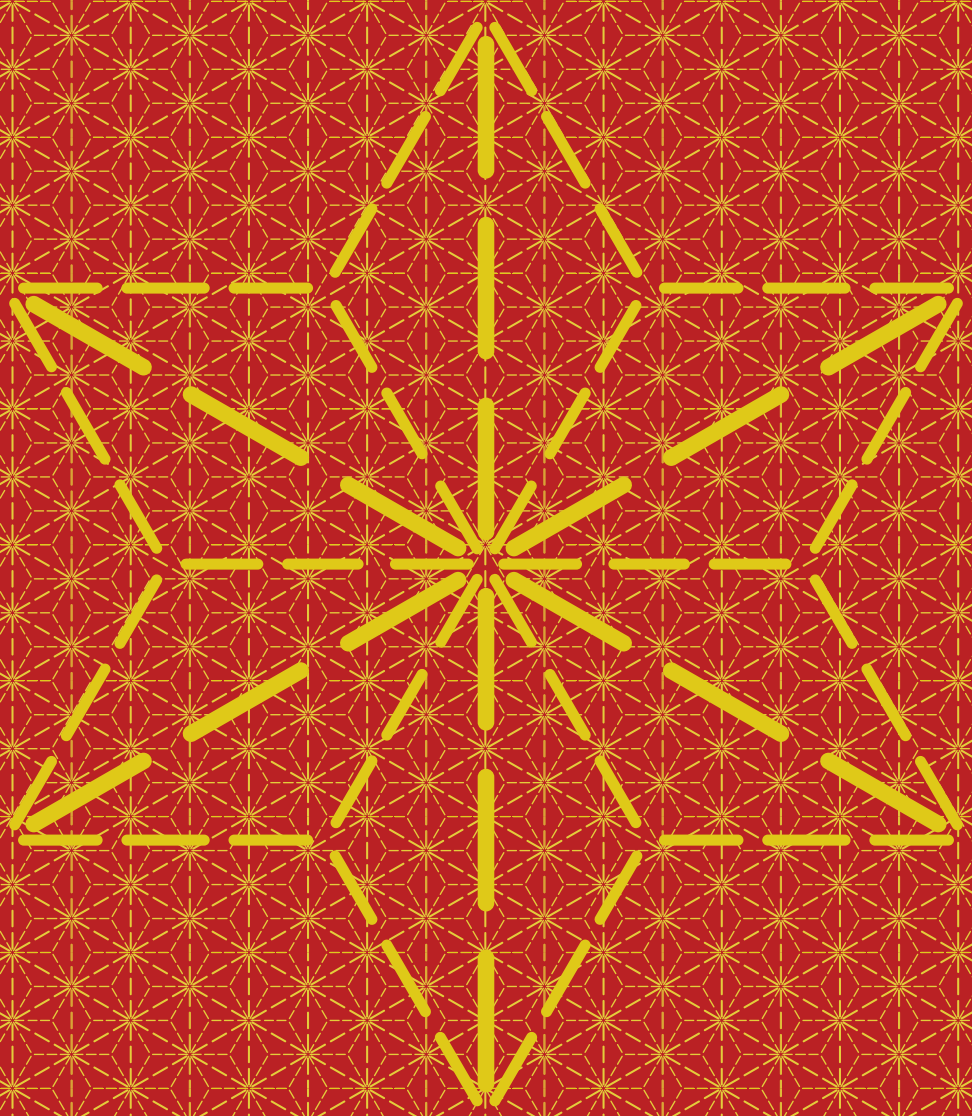


Table of contents

1

Rationale for a course on test development and evaluation

The Test Development and Evaluation Course is offered by the HEC as a professional course in year 4 of the new B.Ed. (Hons) Elementary programme to prepare Student Teachers for effective testing and evaluation practices in primary schools. Student Teachers study this course after learning about the theories related to formative and summative assessment in the Classroom Assessment course in year 2.

Formative assessment has been established as one of the few pedagogical techniques that have a marked positive effect on classroom learning (Black and Wiliam, 2003). In contrast, summative tests can have harmful effects on teaching and learning (Mansell and the Assessment Reform Group, 2009). This course is intended to help teachers avoid the negative impact of summative classroom tests, which are often an external administrative requirement, by creating positive feedback potential so that tests can be used formatively. There are six key steps involved in summative testing: (i) the adoption of agreed standards by the teacher; (ii) the specification of test coverage; (iii) the framing of questions; (iv) marking according to a rubric; (v) analysis of test effectiveness; and (vi) performance feedback for students. These steps are essential in the evaluation of classroom achievement. Testing for other purposes makes different demands; however, these forms of testing are referenced here to help Student Teachers learn what not to do.

Essential knowledge

Any observation on which a judgement is based has an element of error. Fortunately, error is random and does not aggregate: multiple observations repeat the true part of the score but not the error. Therefore, the sum of many observations of the same attribute is mainly true score with very little error.

Students perform better when they know on what basis they will be judged. Teachers should share standards with their classes.

Standards should not merely involve topic coverage, but they should include the kinds of understanding that should be developed. These are defined in taxonomies of objectives, and they are essential guides to the classification of items, questions, and test objectives as a whole.

Reporting raw marks to students inevitably leads to students comparing themselves with each other (norm referencing) on the basis of tiny differences that are wide open to error. Turning raw marks into grades (criterion referencing) focuses attention on standards of subject achievement rather than the performance of peers. Grades are much more inclusive than raw marks. Everyone can succeed instead of a few getting the top ranks and others envying in norm-referenced testing.

Common misconceptions

- *Assessment* and *evaluation* are interchangeable terms and may be considered synonyms.
- Assessment is designed to put students in rank order.
- Test development is not a difficult task.
- Standard setting cannot be applied in classroom examinations.
- The only object of evaluation is student academic achievement.
- Highly reliable tests also possess high test validity.
- Results and statistics used in test analysis do not clarify meaning.

2

Course syllabi

TEST DEVELOPMENT AND EVALUATION

TEST DEVELOPMENT AND EVALUATION

In this section you will find two syllabi that have been written by groups of faculty. Using the HEC Scheme of Studies for the course, they considered the balance between demands of the subject, active learning pedagogies, their students, and the particular university milieu in which they work. The syllabi all reflect the same key concepts and broad goals, but they vary in emphasis.

SYLLABUS 1

By

Dr Bashir Hussain, Dr Muhammad Sarwar, Sadruddin Qutoshi, and Saira Soomro

Year, semester

Year 4, semester 8

Credit value

3 credits

Prerequisite

Successful completion of the Classroom Assessment course

Course description

The Test Development and Evaluation course will focus on knowledge, understanding, and skills in the development of valid, reliable, and adequate tests and evaluation procedures as a means to improve learning. Major topics covered include theories of test development, characteristics of a good test, steps in test construction, alternative assessment strategies, and evaluation and accountability based on value addition. Assessment methods include written tests, assignments, presentations, observation, peer assessment and self-assessment, and portfolios. At the end of the course Student Teachers will be able to plan, organize, develop, administer, and score tests, report student performance, and utilize results to improve student learning.

Course outcomes

After studying the course, Student Teachers will be able to:

- describe and explain types of tests including their advantages and limitations
- differentiate and apply Bloom's and Structure of Observed Learning Outcomes (SOLO) taxonomies for test construction
- describe the role of classical, item response, and generalizability theory in test development
- explain the characteristics of an effective test
- construct tests systematically
- use a variety of essential assessment strategies
- describe and use evaluation to improve learning, teacher performance, and school performance based on value added.

Learning and teaching approaches

There will be a variety of learning and teaching approaches, such as interactive lectures, classroom discussions, presentations, and role play. To equip Student Teachers with testing and evaluation skills, there will be practical exercises in test construction, administration, and analysis. Peer teaching should be used regularly. Active learning strategies, including pair and group work, should also be used regularly.

Instructors and Student Teachers should integrate this course with other courses, including thesis or project work; adapt the course to personal interests, knowledge, experiences, and responsibilities; and design and engage in assignments with sufficient depth and breadth to be useful in other courses and their teaching career.

Student Teachers are expected to explore and apply educational software and other test development and evaluation resources while engaging in a critical examination of educational issues related to testing, assessment, and evaluation.

Unit outlines

Unit 1: Introduction to testing (2 weeks)

This unit aims to provide Student Teachers with an understanding of the term *testing* and its meaning in the context of teaching and learning. Student Teachers will get an overview of testing, different kinds of tests, advantages of testing, and its limitations. The unit will also cover teacher-made tests and standardized tests and how they support teaching and learning. Student Teachers will discuss and analyse the objectives of test development, evaluation, and the implication of evaluation for learners and teachers based on both Bloom's and SOLO taxonomies.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- describe the concept of testing
- explain various types of tests, including their advantages and limitations
- recognize domains of Bloom's Taxonomy in individual questions and items
- explain the use of SOLO Taxonomy for testing.

Unit content

- Concepts of testing
- Testing
- Kinds of tests
- Teacher-made tests
- Standardized tests
- Benefits and limitations of tests
- Concept of taxonomy in testing
- Using Bloom's Taxonomy in test development
- Using SOLO Taxonomy in test development

Unit 2: Characteristics of a good test (2 weeks)

This unit focuses on the characteristics of a 'good test' and outlines its essential elements. It aims to help Student Teachers develop skills in design, trial, moderation, and validation of testing instruments for a range of purposes, including formative and summative assessment and self-assessment. The unit also covers validity and authenticity, reliability, objectivity, adequacy, and differentiability. Components of this unit relate the issues of testing with Student Teachers' examination experiences. This unit is largely practice-based, so in addition to reading texts on testing, Student Teachers will regularly write and participate in classroom discussion.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- explain the characteristics of a good test
- describe and calculate various types of reliability
- ensure validity and adequacy in test construction
- evaluate test items on the basis of their discrimination
- describe the comparability and utility of a test.

Unit content

- Concept of a good test
- Reliability of tests
- Practice session to calculate reliability of tests
- Validity of tests
- Evaluating test items based on their discrimination power
- Utility of a test

Unit 3: Steps of test construction (4 weeks)

This unit focuses on various steps involved in test construction such as planning and organizing the test development process, specification of the test form and procedure, and item analysis. The unit also examines rationales for test development such as specification of learners, learning situation, and test type (e.g. summative, formative, and placement). Student Teachers will develop, administer, and score a test based on specifications and will provide descriptive statistics on test reliability and item analysis. Components of this unit relate the issues of testing to the examination experience of Student Teachers. This unit is largely practice-based.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- ensure content validity (course coverage, concept coverage, learning outcomes coverage) by using a table of specifications
- apply various steps involved in test construction, including constructing tables of specifications based on Bloom's Taxonomy and on SOLO Taxonomy
- construct multiple-choice questions (MCQs) and identify the elements of the questions (i.e. stem, key, and distractors)
- perform item analysis (difficulty, discrimination, fairness)
- score tests while minimizing probable biases
- construct and mark short answer questions
- develop model answers and create marking schemes for essay questions.

Unit content

- Determining the behaviours to be assessed
- Planning the test
- Ensuring content validity (course coverage, concept coverage, learning outcomes coverage) through a table of specifications
- Constructing a table of specifications based on Bloom's Taxonomy
- Constructing a table of specifications based on SOLO Taxonomy
- Writing good MCQs, and constructing tests with MCQs based on a table of specifications
- Reviewing peer's tests and scores
- Performing item analysis (difficulty, discrimination, fairness)
- Constructing short answer questions
    - Marking guides for short answer questions
- Constructing essay questions and tests
    - Developing model answers and marking schemes for essay questions

Unit 4: Essential assessment strategies (5 weeks)

This unit describes the purpose of assessment in teaching the social sciences, how to design rubrics for assessment, and how to align grading policies for summative assessments with the syllabus. This unit also provides guidance on assessing student activities, assignments, and projects, as well as on some innovative methods of assessment. It also will prepare Student Teachers for examinations according to new assessment policies used by the four-year B.Ed. (Hons) Elementary programme in Pakistan.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- explain a variety of essential assessment strategies
- use rubrics to assess students' work (e.g. performance, assignments, and presentations)
- use peer assessment in a complementary way
- describe computer-assisted testing
- use interviews as a tool for assessment.

Unit content

- Classroom observations
  - What is the purpose of classroom observation?
  - Planning and preparing for observation
  - Typical observation formats
  - Deriving results from the observation by developing rubrics
- Assignments and presentations
  - Your intended audience
  - Format, structure, and submission requirements
  - Grading criteria
- Projects
  - Definition of a project
  - Tasks versus tests
  - Five features of a project
  - What makes a project successful?
  - Phases of a project
  - How to assess projects and use them for evaluation
  - Double marking, interrater reliability, and the Spearman–Brown prophecy formula
- Oral questioning
  - Purpose of questioning (e.g. feedback for improving teaching and learning)
  - Guidelines for questioning

- Peer appraisal
  - 'Guess who' techniques
  - Socio-metric techniques
- Interview strengths and weaknesses
  - Interview format
- Portfolio assessment
  - Two types of portfolios (increasing breadth, increasing depth)
  - Steps in the portfolio assessment process
- Computer-assisted testing and the generation of parallel forms for the measurement of change

Unit 5: Evaluation and accountability based on value addition (3 weeks)

This unit will provide some orientation to Student Teachers about evaluating learning, textbooks, and material with rubrics. This unit emphasizes the use of assessment and value-added results for teachers, textbooks, and course evaluation. Moreover, this unit focuses on supporting Student Teachers' role in school accountability.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- explain the concept of evaluation
- use assessment and value-added results for evaluation
- understand how teachers are evaluated on the basis of student performance
- compare teachers based on value-added results in terms of student learning.

Unit content

- Concept of evaluation
  - Using evaluation for different purposes, including teacher and student evaluations
- Accountability and evaluation
- Teacher accountability
- Textbook evaluation
  - Concept of textbook evaluation
  - How to evaluate a textbook
  - What are the basic things to consider in textbook evaluation?
- Concept of course evaluation
  - How and why to evaluate a course
- Designing tools for evaluating teachers, courses, and textbooks
- Review of concepts relating to tests, testing, and evaluation
- Review of the theories of test construction
- Review of concepts relating to evaluation

Grading policy

There are three assessable components in this course.

Mode of assessment	Percentage
Assignments	45%
Midterm exam	25%
Final exam	30%

Assignments may include the following activities:

- critical reflections on readings
- independent learning relating to role of test development and evaluation
- group and individual presentations
- role play and demonstration of technology use
- group tasks
- classroom participation and attendance.

The midterm and final examinations will both be 1.5 hours long and include multiple-choice, short answer, and essay questions. The midterm will cover the first seven weeks, and the final exam will cover the remaining weeks.

Rubrics will be developed in class to assess each assignment.

Recommended reading

Cohen, R., & Swerdlik, M. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7th ed.). Columbus, OH: McGraw-Hill.

Gardner, J. (2006). *Assessment and learning*. Thousand Oaks, California: Sage Publications.

Kline, T. J. B. (2005). Classical test theory: Assumptions, equations, limitations, and item analyses. In T. J. B. Kline (Ed.), *Psychological testing: A practical approach to design and evaluation* (pp. 91–106). Thousand Oaks, CA: Sage Publications. Available from:

➢ http://www.sagepub.com/upm-data/4869_Kline_Chapter_5_Classical_Test_Theory.pdf

Zeng, J., & Wyse, A. (2009). *Introduction to classical test theory*. Lansing, MI: Michigan Department of Education. Retrieved from:

➢ http://michigan.gov/documents/mde/3_Classical_Test_Theory_293437_7.pdf

Miller, M., Linn, R,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

SYLLABUS 2

By

Ishrat Siddiqa Lodhi, Dr Naveed Sultana, Dr Rizwan Akram Rana, Samreen Jalal, Saima Mushtaq, and Zahid Majeed

Year, semester

Year 4, semester 8

Credit value

3 credits

Prerequisite

Successful completion of the Classroom Assessment course

Course description

Awareness of test development, measurement, and evaluation principles and procedures is essential for teachers to monitor students' academic progress. This course is designed for Student Teachers to enhance their level of knowledge, understanding, and practical skills in testing and evaluation. It mainly deals with test development, standard setting, evaluation strategies, grading, and reporting. This course also aims to integrate and implement theory and practice to strengthen the fundamentals of measurement and assessment learned in the Classroom Assessment course. After completing this course, Student Teachers will be in a position to apply the methodology of test development and evaluation in the classroom.

Course outcomes

By the end of the course, Student Teachers will be able to:

- understand the key concepts, methods, and paradigms of test development and evaluation
- apply the key concepts, methods and paradigms of test development and evaluation
- develop, assemble, administer, score, and analyse appropriate tests to interpret and provide feedback on students' progress
- provide balanced assessment aligned with standards and outcomes to improve the teaching-learning process
- apply peer evaluation and self-evaluation tools and techniques (e.g. portfolio assessment, expert evaluation) for feedback purposes
- identify emerging trends in test development and evaluation for future implementation.

Learning and teaching approaches

Teaching approaches

- Lecture and discussion
- Brainstorming
- Self-directed learning and self-study
- Group-based learning (e.g. cooperative learning)
- Literature reviews

Learning tools

- Homework assignments and projects
- Field work, including data collection and analysis
- Maintaining a log or journal

Unit 1: Test development (3 weeks)

A key goal of test development is to create a valid measure of standard-referenced student performance. For this purpose, Student Teachers must know the theoretical principles, issues, and required decisions to develop tests for different purposes. This unit is comprised of three parts: introducing key terms such as *testing* and *measurement*, skills and tools required for test development, and assessment of test tools' credibility.

This unit elaborates on the steps involved in test development. It includes performance assessment, which measures what students can do rather than how much they know, as well as related tasks, which are based on what is most essential in the curriculum and what is interesting to students.

Finally, this unit covers the development of tools and determining the right way to collect data. For example, when designing evaluation tools and selecting evaluation methods, Student Teachers should consider the cultural contexts of the communities in which programmes operate. Overall, this unit aims to impart on Student Teachers the fundamental information, skills, and disposition that will allow them to manage assessment and evaluation programmes at the desired level.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- define *test*, *measurement, assessment,* and *evaluation*
- understand the purpose, principles, and scope of test and evaluation
- describe the concept and process of test development
- comprehend the theory and practice of norm- and criterion-referenced tests
- recognize the role of performance assessment in enhancing the quality of the teaching-learning process
- understand the importance of measuring students' interest, attitude, and creative thinking abilities

- demonstrate a high level of competence in developing and administering different, context-specific evaluation tools
- reduce unnecessary complexity in test items
- overcome language barriers that can threaten the validity of content-based assessment.

Unit content

- Overview of the meaning of *test, testing, measurement, assessment,* and *evaluation*
- Test development process
- Common issues in test development (e.g. language of test, readability, feedback)
- Performance assessment of students
- Developing assessment tools (e.g. rubrics, rating scale, checklist)
- Test administration
- Item analysis

Unit 2: Psychometric properties of formative and summative assessment (2 weeks)

This unit discusses emerging trends in formative and summative assessment. The first section of this unit examines formative and summative assessment as assessments *of* learning and assessments *for* learning.

The second section looks at essential psychometric techniques that provide useful information for the improvement of student learning and assessment procedures. This unit also focuses on improving Student Teachers' competence in classroom assessment in order to yield accurate information about student achievement. It also aims to develop Student Teachers' ability to use the classroom assessment process and its results to enhance learning. Finally, Student Teachers will be provided opportunities to engage in hands-on activities.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- describe the different purposes of formative and summative assessments
- explain formative assessment as a process
- explain summative assessment as a product of learning
- describe the concepts of validity and reliability
- understand evidence of validity and reliability
- explain the threats to the validity of formative and summative assessment
- explain the threats to the reliability of formative and summative assessment.

Unit content

- Relationship between formative and summative assessment
- Test validity
- Test reliability
- Types of reliability

Unit 3: Grading and reporting (3 weeks)

Statistics play a vital role in our daily educational lives. From time to time, a teacher has to collect, organize, and analyse data in order to make decisions about the teaching-learning process. This unit will provide information about basic descriptive statistics (measures of central tendency and measures of variability) so that the Student Teachers can analyse measurements and present conclusions. It aims to help Student Teachers understand the functions of grading and different types of grading procedures. This unit will enable Student Teachers to assign grades to students by using the most effective grading practices, criteria, and standards to provide accurate, specific, and timely feedback that can help improve student performance. Additionally, this unit will offer hands-on experience in interpreting test scores and reporting student performance.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- recognize the basic symbols of elementary statistics
- compute measures of central tendency and variability
- analyse the results of measurements
- comprehend the function, types, and uses of grades
- apply principles and strategies and establish criteria to make grading efficient, consistent, and fair
- communicate learning expectations to students by establishing grading standards
- interpret test results with necessary caution and describe problems in grading.

Unit content

- Elementary statistics
- Effective grading in the classroom
- Establishing criteria and standards for grading
- Interpreting test scores
- Reporting assessment results to students, teachers, parents, and school administration
- Problems in grading and reporting

Unit 4: Curriculum and test development (3 weeks)

This unit is focused on the relationship between curriculum and test development. In this unit, Student Teachers will develop an understanding of how learning outcomes are used to create achievement tests. It also provides an overview of the National Professional Standards for Teachers of Pakistan with special reference to Standard-05, which is based on assessment and curriculum. This unit will examine the role of curriculum in test development and assessment, and discuss the role of teacher's expectations of students' performance. It will also throw light on how a teacher can give feedback directly related to student performance that supports, rather than hinders, student potential.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- understand the relationship between curriculum and the process of test development
- analyse the best strategies for sharing and communicating assessment results
- communicate positive expectations to their students effectively.

Unit content

- Norm-referenced and criterion-referenced tests
- Linking curriculum with test development
- Communicating and sharing standards (National Professional Standards for Teachers and National Curriculum 2006)
- Effect of teacher expectations on student achievement

Unit 5: Outcomes of assessment results (3 weeks)

This unit aims to enable Student Teachers to assess student progress by improving teaching and learning. There are a number of formats, tools, and methods available to assess student progress, but in Pakistan, teachers are bound to follow the format provided by the provincial Directorate of Education. This unit will provide a taste of other formats to which adaptations can be made to better suit the local educational context. Assessment results are not only important for students and teachers but also for school effectiveness, improvement, and development. The overall focus of the unit is to train Student Teachers to assess students' progress, develop students' portfolios, make decisions in terms of instructional purposes, and use assessment results for school effectiveness and improvement.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- understand the concept of student progress
- develop student portfolios and evaluate them
- make instructional decisions by assessing student progress
- apply instructional decisions for effectiveness and school improvement.

Unit content

- Student progress
- Decision-making for instructional purposes
- School effectiveness and classroom improvement

Unit 6: Emerging trends (2 weeks)

This unit covers emerging assessment techniques, models, and approaches that will enhance Student Teachers' understanding of assessment in the current educational climate. Self-assessment and peer assessment are new trends in educational assessment. International assessments, including the Trends in International Mathematics and Science Study (TIMMS) and the Programme for International Student Assessment (PISA), and national assessments, such as the National Education Assessment System (NEAS) and the Annual Status of Education Report (ASER), are different from classroom tests and school examinations. The information is aimed to equip Student Teachers with the knowledge, understanding, and skills to adapt some emerging trends in assessment for their future primary school classes.

Unit outcomes

After completing this unit, Student Teachers will be able to:

- predict their position relative to their past learning
- predict their relative position with respect to their peers
- differentiate between self-assessment and peer assessment
- compare and contrast assessments by international and national assessment agencies.

Unit content

- Measuring student growth
- Self-assessment, peer assessment, and reducing the burden on teachers
- International and national assessment agencies – NEAS, AKU-EB, and Federal Board of Intermediate and Secondary Education (FBISE)

Course assignments

Suggested course assignments involve the development and administration of test items with subsequent item analysis and report preparation. Student Teachers will also be asked to develop a portfolio.

Grading policy

The grading policy depends upon the individual university. For example, the University of the Punjab, Lahore, follows the following pattern for its four-year undergraduate programmes:

Mode of assessment	Percentage
Assignments	25%
Midterm exam	35%
Final exam	40%

Recommended books

Alastair, I. (2007). *Enhancing learning through formative assessment and feedback*. London: Routledge.

Banks, R. S. (2005). *Classroom assessment: Issues and practices*. Upper Saddle River, NJ: Pearson Education.

Black, P., Harrison, C., Marshell, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Milton Keynes, UK: Open University Press.

Ebel, R. L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. New Delhi: Prentice Hall of India.

3

Representative syllabi with teaching notes

This section contains an example syllabus with accompanying teaching notes. The teaching notes have been integrated around broad themes addressed in the course (test development, reporting, and evaluation) and offer opportunities for hands-on practice to prepare Student Teachers for effective testing and evaluation in their careers. Faculty who are teaching the course for the first time or who are interested in the process of curriculum design may find it useful to see how the authors of this representative syllabus chose to develop particular ideas and themes in their notes. Ideas presented in teaching notes here are not duplicated in the later section where integrated themes may be found.

SYLLABUS

By

Dr Khalid Saleem, Naila Siddiqua, Raqeeb Imtiaz, and Dr Saifullah

Year, semester

Year 4, semester 8

Credit value

3 credits

Prerequisite

Successful completion of the Classroom Assessment course

Course description

The present course covers test development and evaluation at the B.Ed. (Hons) level. This course will enable Student Teachers to select assessment standards relevant to the content area they are teaching from the National Curriculum 2006 and the National Professional Standards for Teachers. It will also provide hands-on activities to help Student Teachers become proficient in test development and reporting. This course will also focus on item analysis, standardized test development, and basic statistical techniques for judging a test's reliability and validity. Student Teachers will have the opportunity to examine different perspectives on student, teacher, and textbook evaluation.

The pedagogy for this course will include interactivity, hands-on work in different content areas, skill-based and performance-oriented activities, presentations, observation, and practicums involving the main stages of test development and interpretation.

Course outcomes

After completing this course, Student Teachers will be able to:

- develop student achievement tests
- comprehend and apply standard-setting techniques
- apply taxonomies of educational objectives in test development and assessment
- analyse test items, including item difficulty level and discrimination
- apply basic statistics for reporting test results
- convert raw scores into standard scores.

Learning and teaching approaches

- Lectures followed by discussion using learner-centred methods
- Cooperative learning
- Assignments involving practical tasks in test construction and analysis
- Quizzes
- Projects using learner-centred or do-it-yourself methods

Unit outlines

Unit 1: Test development process (3 weeks)

This unit sets out the test development process's conceptual framework within the context of the National Curriculum. It presents information on tables of specification as well as item and test forms. It also discusses instructions to examinees, answer keys and rubrics, and examination and feedback report.

Unit outcomes

At the end of this unit, Student Teachers will be able to:

- understand the test development process
- apply taxonomies of educational objectives in test evaluation
- construct items, test forms, examination reports, and other tools used in test development.

Unit content

- Concept of test
- Defining test objectives
- Designing tables of specification
- Constructing test items
- Test form development
- Test conduct
- Marking and result compilation
- Writing the examination report
- Writing the feedback report

Unit 2: Concept of standard setting (3 weeks)

This unit focuses on standards as a concept as well the different types of standards and how they are set. In particular, two methods of setting performance standards are examined: item-centred and examinee-centred methods.

Unit outcomes

At the end of this unit, Student Teachers will be able to do the following:

- define the concept of standards in educational and testing contexts
- differentiate between content- and performance-based standards
- identify the need for standard setting in assessment
- comprehend and apply standard-setting methods in the assessment process.

Unit content

- Use of standards in education
- Use of standards in testing
  - Need for standard setting
  - Content-based standard setting
  - Performance-based standard setting
- Methods of setting performance standards
  - Item-centred methods
  - Examinee-centred methods

Unit 3: Interpreting testing and evaluation scores using statistics (4 weeks)

This unit discusses the role of statistics in assessment and evaluation. Student Teachers will learn to summarize data, relate it to the population, and present data using different graphs. This unit explores the sources of variation and the relationship between two variables. It also analyses tests with respect to reliability and validity.

Unit outcomes

At the end of this unit, Student Teachers will be able to do the following:

- recognize the role of statistics in assessment and evaluation
- display data graphically
- understand the central tendency and variation in data
- measure the relationship between two variables
- comprehend the relationship between reliability and validity.

Unit content

- Introduction to statistics
- Frequency distribution
- Types of graphs
- Measures of central tendency
- Measures of dispersion

- Measures of relationship
- Item analysis
  - Ensuring reliability and validity

Unit 4: Reporting the results of standardized tests (4 weeks)

This unit enables Student Teachers to interpret commonly reported scores such as percentile ranks, percentile band scores, standard scores, and grade equivalents.

Unit outcomes

At the end of this unit, Student Teachers will be able to do the following:

- understand the different frames of reference to interpret scores
- comprehend different types of standardized scores
- transform scores from one scale to another
- examine and interpret published norms.

Unit content

- Methods of interpreting test scores
  - Criterion-referenced interpretations
  - Norm-referenced interpretations
- Grade equivalent scores
  - Percentile rank
  - Stanines
  - Calculating standard scores
- Standard scores
  - z-score
  - T-score
- Normalized scores
  - Calculating normalized scores
- Criteria for judging norms
- Cautions in interpreting transformed test scores

Unit 5: Evaluation context (2 weeks)

This unit focuses the three basic elements of the educational process: students, teachers, and textbooks. It aims to help Student Teachers understand the different parameters used to evaluate teaching-learning skills and competencies. It also examines the various components of a textbook and how to judge a textbook's quality. Student Teachers will have the opportunity to perform a hands-on activity using different tools for evaluation.

Unit outcomes

At the end of this unit, Student Teachers will be able to do the following:

- understand the differences in evaluating different elements of a school (e.g. teachers, student learning, textbooks, and school environment)
- identify indicators to evaluate different elements of school
- apply different tools and checklists for evaluation.

Unit content

- Student evaluation
- Teacher evaluation
- Textbook evaluation
- Project work involving tools of evaluation

Course assignments

This course contains two assignments.

Developing an achievement test

Student Teachers will maintain a file containing a table of specification, items developed and a related mark scheme and item classification table, a first draft of a test with comments from peers, a printed copy of the test form, a compiled result sheet from one class, and two reports. It is recommended that this assignment accounts for 20 per cent of the class mark.

Performance standard-setting exercise

Student Teachers will complete a report on performance standard-setting exercise they conducted. The report will include the procedure adopted and recommendations for a method of setting performance standards. It is recommended that this assignment accounts for 10 per cent of the class mark.

Grading policy

A variety of assessments will be used in the course, including midterm and final examinations as per relevant university rules.

Recommended resources

6 misconceptions about standardized tests (2012). Retrieved from:

- http://www.minglebox.com/article/sat-exam/6-misconceptions-about-standardized-tests

Ebel, R. B., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Haladyna, T. M. (1999). *Developing and validating multiple choice test items* (2nd ed.). Mahwah, NJ: Erbaum.

Roid, G. H., & Haladyna, T. H. (1982). *A technology of test item writing*. London: Academic Press, Inc.

Shrock, S. A., & Coscarelli, W. C. (2007). *Criterion-referenced test development* (3rd ed.). San Francisco: Pfeiffer.

Van der Linden, W. J. (2003). *Linear models for optimal test design: Statistics for social science and behavioural sciences.* New York City: Springer.

TEACHING NOTES

Unit 4: Reporting the results of standardized tests (4 weeks)

Developed by Naila Siddiqua

Methods of interpreting test scores

Criterion-referenced interpretations

Student Teachers will think, pair, share to explore the advantages and disadvantages of relating classroom scores to national standards.

Norm-referenced interpretations

Student Teachers will conduct a three-step interview to explore the ranking of students in a class and who benefits.

Grade equivalent scores

Student Teachers will engage in data calculations to compare results.

Percentile

A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the score below which 20 percent of the scores may be found; if a score is in the 86th percentile, it is higher than 86% of the other scores.

Stanines

Percentile scores have the disadvantage that they can give a false impression of test accuracy. If there are more than 100 testees, we may see every reported percentile from 0 to 99 on a test that has only 50 marks. Stanines get round this difficulty by reporting just nine outcomes – hence, the name *standard nine*. Stanines consider the testee's position in the rank order, but the rank order is related to the normal curve.

The following website provides more information about stanines:

- http://assessment.tki.org.nz/Using-evidence-for-learning/Concepts/Concept/Percentages-percentiles-and-stanines

Calculating standard scores

Standard scores are based on the standard deviation. A standard score compares a testee's score to the average score by reporting how many standard deviations separate the testee's score from the mean average. A z-score is the difference between the score and the mean, divided by the standard deviation. z-scores almost always fractions that vary between +3.0 and –3.0, so they are not well suited for reports intended for parents or anyone else who is not familiar with statistics. An alternative is the T-score, which converts the observed mean score to 50 and the observed standard deviation to 10, resulting in a reported score range between 20 and 80. If you plot a distribution graphically, it will usually have more observations towards the middle score and fewer

at the extremes. However, it can be symmetrical, with high scores bunched and low scores spread out. This bunching reduces scores' correlation with any other variable that does not have the same distribution. As such, when the relationships between scores on different tests are of interest, the scores can be normalized (i.e. given the same normal distribution). Raw scores are ranked and turned into percentiles. An equivalent z-score is read off a table of the area under the normal curve without any further calculation.

Standard scores

Student Teachers will engage in data entry and calculation to compare the results. This comparison will be made by converting raw scores into standardized scores. Student Teachers can use the data available at http://www.udel.edu/educ/gottfredson/451/Glutting-guide.htm

z-score

Student Teachers will engage in calculations to convert raw scores into standardized scores.

T-scores

Student Teachers will engage in calculations to convert raw scores into standardized scores.

Normalized scores

Student Teachers will engage in calculations to convert raw scores into normalized scores.

Unit 5: Evaluation context

Suggestions for Student Teacher evaluation

Student Teachers will keep a reflective journal for three weeks in which they consider how their performance in the course is developing. They will then compare notes. What have they learned? Are they all learning at the same rate? How fast should they be learning?

Student Teachers will research tools to evaluate students. They will create a presentation (e.g. PowerPoint presentation) that covers at least three tools, such as observation, checklist, rating scale, and anecdotal records.

Based on their research, Student Teachers will then develop tools for student evaluation. For example, they could develop a checklist.

Suggestions for teacher evaluations

Have Student Teachers work to define teacher effectiveness. Have them engage in a think, pair, share to identify the criteria for teacher effectiveness. Consider providing them with 'Teacher evaluation' (or the 'Design standards' section), available from:

- http://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf?files/Teacher

Have Student Teachers investigate methods for teacher evaluation. Ask them to read and discuss articles on principles and models of effective teaching. Potential reading material includes the following:

- Guidelines for evaluating teaching:
  - http://www.crlt.umich.edu/tstrategies/guidelines
- Principles of effective teaching and examples of descriptors:
  - http://www.brookline.k12.ma.us/index.php?option=com_docman&task=doc_download&gid=147

Tools for teacher evaluation: Ask Student Teachers to modify the teacher evaluation tools to fit with local expectations.

Ask Student Teachers to develop a teacher evaluation checklist for different subjects.

As homework, Student Teachers will write a paragraph about a method they believe is suitable for teacher evaluation. They should explain why they think this method is appropriate.

Student Teachers should try to use the checklist they developed and/or the method they selected at a practicum school. They can then present their findings and reflections on using the checklist or other method in seminar.

4

Integrated teaching notes

During the curriculum development process, contributing faculty were encouraged to keep notes that would be useful to others who may teach this course in the future. These were submitted along with the course syllabi. Teaching notes include ways to introduce the course, ideas for teaching units and sessions, sample lessons plans, and suggestions for reading and resource material. These have been integrated into a single section of this document to create a rich and varied collection of ideas that is easily accessible to others. Except in cases where there is duplication of ideas, faculty are credited for their contribution.

INTRODUCING THE COURSE

Key words

- Test
- History of testing in education
- Test objectives
- Table of specification
- Test items
- Test forms
- Test conduct
- Marking
- Examination report
- Feedback report

Introducing tests as a concept

Developed by Isbah Mustafa

Introduction

Ask Student Teachers the following questions to gauge their experience with and knowledge of tests:

- Which tool did this institution use to decide your admission to the B.Ed. (Hons) Elementary programme? What tool was used to determine whether you successfully completed a semester or year?
- What are the characteristics of a test?

As Student Teachers respond to the questions, make sure that the following points are made:

- Tests are one of the most commonly used assessment tools to evaluate a learner's prior knowledge and skills in a specific field.
- A test is a collection of questions, items, and problems designed to determine a testee's knowledge, comprehension, or ability.

History of testing in education

Use the 'Brief history of testing in education' handout to introduce the history of tests in education. Imperial examinations and the Stanford–Binet test represent major developments in educational testing. Introduce these developments and then discuss tests used in Pakistan in the field of education, such as certificate examinations, diagnostic tests by NEAS and Provincial Education Assessment System (PEAS), and university admission tests. These tests are used for different purposes such as selection, promotion, placement, and aptitude identification.

Handout

Brief history of testing in education

S

The imperial examination

This examination is considered to be the first selection test. It was used in Imperial China to select young officers for state's bureaucracy. It dates to the Sui Dynasty (581–618). Before this test, the selection of bureaucrats was made through personal references. This first objective tool for selection on merit had two defining features: it was timed (24 hours) and it had set questions that were given to all candidates. However, its scope was limited in terms of the variety of skills and content knowledge to be tested. Nevertheless, the test offered a way to objectively screen youth for high posts and entry into the elite class.

Until the 19th century, educational achievement was assessed only in universities, but this changed with the emergence of the movement for mass/compulsory education and the introduction of school accountability. In 1863 in England, a public finance system, 'payment by results', was introduced that made national funding of individual schools dependent in part on student performance on standard examinations, usually oral questioning by visiting school inspectors. The system collapsed in 1890; the increased number of schools and teacher unionization made the system cumbersome and generated bureaucratic paperwork.

A performance standard

The examination standards used in Imperial China and 19th-century England were combined in the US Army Aptitude tests, Alpha and Beta, which were the first vocational aptitude tests designed for large-scale group administration. They used all multiple-choice items, and the Beta version was the first non-verbal test that allowed speakers of languages other than English to be fairly assessed.

Binet–Simon intelligence test

Alfred Binet and Théodore Simon developed the first successful intelligence test in 1905. This one-to-one test was meant to identify students requiring more input in French schools. Based on student performance, the test reported testees' performance relative to their age. Thus, the Binet–Simon intelligence test was born.

Stanford–Binet intelligence test

The Stanford–Binet intelligence test is a modified version of the Binet–Simon intelligence test. This test is widely used to measure intelligence quotient (IQ) rather than mental age. It uses a scale ranging from below 20 to above 145 to reflect the testee's gauged intelligence. It was first presented in 1916 and has been revised since, most recently in 2003. The Stanford–Binet test is currently used for neuropsychological and career assessment, placement, compensation evaluations, forensics, and aptitude.

Current test practices in Pakistan

Have Student Teachers work in groups of four to six. Ask each group to identify a type of test that most of the members have come across in their academic careers (e.g. promotion tests, aptitude tests, or selection tests). Alternatively, they could reflect on the test they developed in the Classroom Assessment course in semester 4. The group should discuss the following features of the test:

- what it intended to measure
- what it actually measured
- test structure including length, maximum marks, and question types
- the test's development procedure including the steps, timeline, number of people involved in the procedure, their expertise, and effort required form the testee.

Each group will present a summary of their discussion to the class.

Closing

Summarize the group presentations by listing the tests and finding commonalities, if any, with the imperial examination, Binet–Simon test, or Stanford–Binet test.

References

Imperial examination. (n.d.). Retrieved 27 June 2013 from:
➢ http://en.wikipedia.org/wiki/Imperial_examination

Stanford–Binet intelligence scales (n.d.). Retrieved 27 June 2013 from:
➢ http://en.wikipedia.org/wiki/Stanford%E2%80%93Binet_Intelligence_Scales

Worsfold, A. (n.d.). A history of school examinations. Retrieved from:
➢ http://www.change.freeuk.com/learning/howteach/exams.html

Defining test objectives

Developed by Isbah Mustafa

Introduction

Review the discussion from the previous session regarding the various abilities that tests measure, including specific skills, attitudes, cognitive skills, and knowledge. These target abilities are the objectives of a test, and they underlie the purpose of any test. For example, the objectives of a university admissions test gauge the skills, knowledge, and aptitude necessary to become a successful student once enrolled in a programme. In contrast, the objectives of a promotion test gauge a candidate's skills and knowledge attained beforehand.

Test objectives ensure that tests are aligned with target cognitive levels, difficulty levels, or any other measurements of student abilities. Objectives also help determine the length of the test, as they inform the breadth and depth of the domains covered by the test. Characteristics of good test objectives include consideration of learner qualities, targeted learner behaviour or competency, target content area, and test conditions.

Achievement tests

Introduce the concept of achievement tests, which are the most common tests administered in schools. (In Pakistan, an achievement test is also known as a promotion test.) The objectives of an achievement test are specific to particular content areas, skills, and cognitive levels. Test objectives are driven by the purpose of a test—for example, achievement tests measure student attainment of knowledge and skills in a particular subject or grade.

Ask Student Teachers to think of a subject that was covered by an achievement test they took in primary school such as English or maths. They should consider what those tests were designed to measure (e.g. reading skills), including particular learning outcomes (e.g. the ability to identify the main idea in a text).

The objectives of an achievement test are the learning outcomes that students are expected to achieve. The National Curriculum 2006 refers to these outcomes as Student Learning Outcomes (SLOs). The learning outcomes are available for review from:

- http://passpunjabedu.com/index.php?option=com_phocadownload&view=category&id=4:curriculums-2006&Itemid=34.

Selecting test objectives

Distribute the handout 'Selected class 6 SLOs in English, mathematics, and science', which covers SLOs for class 6. (The content is from the National Curriculum 2006.)

Have Student Teachers work in groups of six. Ask them to think about an achievement test in class 6 English, science, or maths. The group will select SLOs in their chosen subject and create a 30-minute test. They should consider the specific topic or skill that each SLO addresses.

Peer review

This activity is a preamble to an exercise that will run throughout the unit on test development. The groups will conduct a peer review of the selected objectives and answer the following questions:

- Do the selected SLOs provide sufficient information to measure achievement?
- Are the selected SLOs measureable through a paper–pencil test?

Groups will provide the feedback to each other and revise the selection of SLOs as needed.

Handout

Selected class 6 SLOs in English, mathematics, and science

	English	Science	Mathematics
1	Scan a simple text for specific information	Compare the structure of a monocot leaf and a dicot leaf in terms of shape and venation	Read numbers up to 1 000 000 000 (one billion) in numerals and in words
2	Make simple inferences using the context of the text and prior knowledge	List the function of cotyledon	Add numbers of complexity and of arbitrary size
3	Express understanding of a story through role play	Describe different kinds of pollution	Divide numbers, up to six digits, by a two-digit and three-digit number
4	Locate specific information in a calendar, a class timetable, and a report card	Describe the properties of three states of matter on the basis of the arrangement of particles	Find highest common factor of three two-digit numbers using: • prime factorization method • division method
5	Write a story using the elements of story writing	Describe friction and its causes	Multiply a fraction by a number and demonstrate multiplication with help of diagrams
6	Write short informal invitations for a variety of purposes to demonstrate the use of conventions of short invitations	Explain gravitational force using different examples	Divide a decimal with a whole number
7	Identify and use appropriate tone and non-verbal cues for different communicative functions	Investigate light that travels in a straight line	Solve real-life problems involving conversion, addition, and subtraction of units of distance
8	Pronounce and spell more words with silent letters such as *tch* in *switch*, *sch* in *school*	Describe the flow of an electric current in an electric circuit	Use protractor to construct: • a right angle • a straight angle • reflex angles of different measure
9	Locate, provide, connect and use words similar and opposite in meaning	Describe the effect of moisture on soil characteristics	Read a simple bar graph given in horizontal and vertical form

Source: *National Curriculum 2006 for English, mathematics, and science.*

Designing a table of specification

Developed by Isbah Mustafa

A table of specification

A table of specification is the blueprint for a test. It specifies the content coverage, question types, marks distribution, and sometimes cognitive levels.

Engage the class in a recall exercise by asking the Student Teachers about different levels of Bloom's Taxonomy and their description.

Developing a table of specification for an achievement test

Have Student Teachers work in the same groups of six as they did in the selecting test objectives activity (involving the handout 'Selected class 6 SLOs in English, mathematics, and science'). Explain the structure of the table of specification in the handout 'Table of specification', copies of which should be distributed to the class. Each group will develop a table based on the SLOs they selected earlier.

One member of each group will present the table of specification to the class. A member of the group will need to take notes on the observations and feedback offered by the class and the Instructor.

Each group will revise their table of specification based on this feedback.

Template for a table of specification

	Number of Questions and marks to question type		Number of questions and marks to cognitive level		
Topic/Skill	**Close-ended questions**	**Open-ended questions**	**Knowledge**	**Understanding**	**Application**

Constructing test items

Developed by Dr Thomas Christie and Isbah Mustafa

Introduction

Initiate a discussion by asking the class what qualities a good question has. Make sure the following points are made about good questions during the discussion:

- They precisely measure the targeted outcome.
- They have a refined structure, language, and rubric, which reinforce precision.

Discuss these qualities in relation to how they apply to open- and close-ended questions.

Related information is available from the following resources:

- Withers, G. (2005). *Module 5: Item writing for tests and examinations*. Paris: International Institute for Educational Planning/UNESCO. Retrieved from:
  - http://www.iiep.unesco.org/fileadmin/user_upload/Cap_Dev_Training/Training_Materials/Quality/Qu_Mod5.pdf
- Item writing and review guide (2011). Retrieved from:
  - http://www.assess.com/docs/ASC_Item_Writing.pdf

What is a mark scheme (or rubric)?

Introduce the concept of a mark scheme, which is part and parcel of a question, although it does not appear in the question itself. A mark scheme contains information about the maximum marks and their distribution, content area or skill covered by a question, a question's difficulty level or cognitive level, possible answers or areas to be covered in the response, and a stimulus and its source. The mark scheme for questions for the paper specification developed earlier will also capture the SLOs on which a question is based.

Command words

Give Student Teachers a copy of the handout 'Command words'. This activity will help Student Teachers recall the command words associated with Bloom's Taxonomy. These command words are a very effective tool in determining the cognitive level a question is trying to measure.

Ask Student Teachers to recall the command words and the level to which they belong. Record the command words and their respective cognitive level on the board. Note that some command words fall in more than one category.

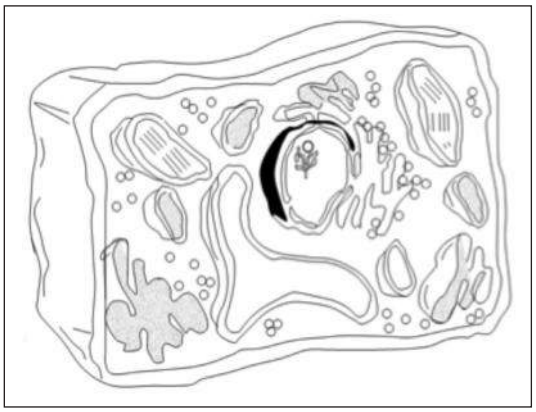
The following websites offer lists of command words:

- Bloom's Taxonomy action verbs (n.d.). Retrieved from:
  - http://www.clemson.edu/assessment/assessmentpractices/referencematerials/documents/Blooms%20Taxonomy%20Action%20Verbs.pdf
- Maynard, J. (n.d.). Bloom's Taxonomy's model questions and key words. Retrieved from:
  - http://www.cbv.ns.ca/sstudies/links/learn/1414.html
- Bloom's Taxonomy with key words (n.d.). Retrieved from:
  - http://achieve.psdr3.org/bloom.html
- Have Student Teachers review and identify the cognitive level in Bloom's Taxonomy that is being tested.

Handout

Command words

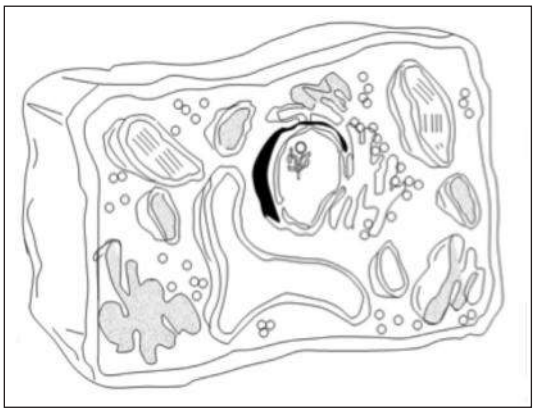
1.Label the picture of the plant cell.



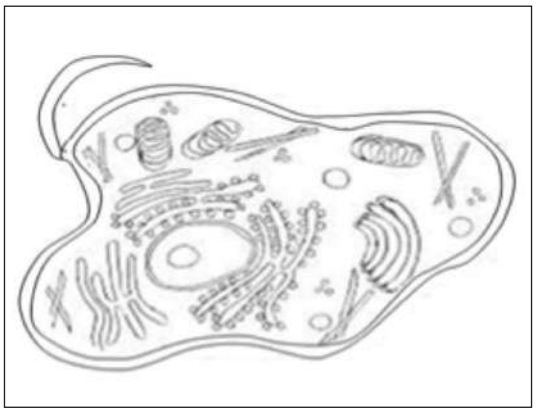
Plant cell

2.Explain a plant cell.

3.Decide which of the two cell pictures (below) is of a plant cell. Justify your selection by listing two characteristics of a plant cell that can be identified in the selected picture.



Cell picture 1



Cell picture 2

Difficulty level

Introduce difficulty level. Explain that it defines the challenge the question poses to the testee. There are different methods of measuring difficulty, including the following:

- calculating the proportion of testees who respond correctly to the question
- estimating difficulty before the question is used.

With the latter method, each question is categorized as easy, moderate, or difficult based on the estimated percentage of students who will respond correctly to it.

In modern development, item response theory (IRT) (Rasch, 1960), is a 'test analysis procedure that assumes a mathematical model for the probability that an examinee will respond correctly to a specific test' (from http://www.nationsreportcard.gov/glossary.asp).

In IRT, item difficulty and candidate ability are reported on the same scale, which is distributed like a normal curve. Item difficulty indicates what level of candidate achievement is associated with a 50/50 chance of success on the item.

Evaluating test items on the basis of their discrimination power

The discrimination of an item is calculated to make sure that it is contributing to the reliability of the test, or its internal consistency. The idea is that those who achieve high scores on an entire test are more likely to get a given item correct than those who are not high scorers. The calculation is straight forward:

1. Rank the total scores.

2. Divide the scores into three groups: top, middle, and bottom. If they do not divide evenly into three groups and there is a remainder, assign the remainder(s) to the top and bottom groups in equal numbers. For example, if you have 107 scores, each of the three groups will get 35 scores with two scores remaining. In this way the items are ranked appropriately.

3. For each item, count the number of testees in each group who got the item correct.

4. Subtract the number of correct testees in the bottom group from the number in the top group, and divide the difference by the number in the bottom group.

5. You now have a proportion, called E_{13}, between +1.0 and –1.0.

An item where $E_{13} = +1.0$ is ideal. It indicates precisely what the test item measures.

Items where E_{13} is zero or negative should not be used. Unfortunately, rewriting the stem or answer options rarely works. It is better to think of a new item from the same cell of your specification grid. If that does not work, reconsider your specification grid. Does that cell or topic really belong with the rest of the test?

Note that in the Rasch model, every item is assumed to have the same discrimination. However, once you analyse a test, it will become very clear that reality does not always reflect the model.

The task of question writing

Ask Student Teachers to work with the same group of six from earlier lesson. Each group will be divided into three pairs. For homework, each pair must develop a set of questions for class 6 students based on the table of specification developed in a previous lesson. The answer key for close-ended questions, the mark distribution, and possible answers for open-ended questions should be included in the mark scheme. A summary of the mark scheme should be submitted in the 'Item classification table' (template below).

Template: Item classification table

Item no.	Marks	Question type	Topic	Cognitive level	Estimated difficulty level

Test development

Developed by Isbah Mustafa

Activity on test development

Have Student Teachers work in their groups of six from previous sessions. Ask each group to develop a test that should take class 6 students 30 minutes to complete. This task involves:

- selecting items from a pool of items that has three times as many items as needed
- designing the front page of the test, which should include instructions, an explanation of the marks, the allocated time, and the types of questions
- matching the pool of selected questions with the table of specification
- grouping the questions in the form by topic and the topics in ascending order according to difficulty level.

Peer review

Ask the groups to peer review each other's tests as homework. They should use the 'Review checklist' handout for this task. They should also offer notes and comments to provide additional feedback.

Based on the feedback received, each group should prepare a final draft of their test.

Handout

Review checklist

No.	Checking points	Tick if the answer is yes
1	Are the instructions on the title page correct and complete?	
2	Is the formatting uniform?	
3	Are the marks listed correctly for each question?	
4	Is the number sequence correct for the questions?	
5	Do the total marks and question numbers match with the instructions on the front page?	
6	Is the paper free of grammatical or spelling errors?	
7	Have all the topics in the table of specification been covered?	
8	Has the weight given to the cognitive levels in the table of specification been followed?	
9	Will the students be able to finish the test in the time allotted (i.e. 30 minutes)?	

Adapted from a checklist developed by AKU-EB for Institute for Business Administration Sukkur, standardized tests for classes 5 and 8, prepared by Isbah Mustafa.

Test conduct and marking

Developed by Isbah Mustafa

Test conduct

Each group must arrange visits to three classes of approximately 40 class 6 students to administer their sample test. The visit should be scheduled for 90 minutes, and Student Teachers will go in pairs.

The pairs will introduce themselves to the students, explain the purpose of the test, and ask for volunteers to take it. The students will have 30 minutes to complete the test. The pairs will provide the required instructions before, during, and after the test and will ensure the availability of answer booklets, printed question paper, or other essential stationery.

Marking responses

After the test has been completed, the Student Teachers will ask the student volunteers to exchange test papers or answer booklets so that they can mark each other's papers (peer marking). One member of each Student Teacher pair will share the correct answers and the other one will roam around the room to assist students. There should be a couple of open-ended questions that require the marker's judgement. Advise the student volunteers on the criteria they should use to mark these responses. Collect the marked tests to complete the 'Result sheet' handout. (Student Teachers should retain the completed tests for future activities.)

Homework: Result sheet development

Each group will compile the results for each class by adding the number of students responding correctly. A compilation sheet that includes all three classes should also be completed. The template for the 'Result sheet' should be used to complete the homework.

Template: Result sheet

Result sheet					
School:			**Date of exam conduct:**		
Examiners:			**Total students in the class:**		
Item no.	**Topic**	**Number of students responding correctly**	**Percentage of students responding correctly**	**Actual difficulty level**	**Estimated difficulty level**

Easy: 70% or more students responded correctly; moderate: 40–70% students responded correctly; difficult: 20–40% students responded correctly.

Writing the examination report

Developed by Isbah Mustafa

Discussion on examination reports

Student Teachers will discuss the result sheets in their groups of six. Initiate the discussion by posing the following questions to the class:

- Which items have mismatched estimated and actual difficulty levels? Why did the mismatch likely occur?
- On which SLOs have students performed well? Poorly?

Each group will address these questions. One group member should take notes on the discussion for writing the examination report.

Writing an examination report

Each group will write an examination report. The report's intended audience is the Instructor, and its purpose is to convey information about the test in general and the challenges emerging based on the questions regarding difficulty level and student performance. Discuss the specifics of the text, including the subject, the testees' characteristics, and the overall results. Issues relating to the difficulty level and student performance should each have their own section.

Writing a feedback report

Developed by Isbah Mustafa

Discussion on examination report

Student Teachers will work in their groups of six and discuss examination reports. It will be necessary for them to use the compiled result sheet and the individual sheets for each class.

Begin by asking them which class performed the best and worst. Also, ask if there was any consistency in class performance for each topic.

Each group should discuss class features that may have created differences. They should also think of ways to help classes to overcome low performance, such as getting stronger students to work with the weaker students or re-teaching the more troublesome topics.

The groups will also consider the graphs and tables to be included in the report. One group member should take notes of the discussion for writing the report.

Sharing notes for the feedback report

A member from each group will share the group's notes with the class. The presentation should address way to help low performing students and classes, and explain how the data will be presented in the report. The rest of the class will give feedback, which should be recorded for use.

Homework

Each group will write feedback reports for each class that took their sample exam. The report should address class performance and student performance. Simple graphs and tables should be used to present information, particularly to examine student performance. Also, suggest remedial action for classes with low performance in a given content area.

Unit assessment

Student Teachers should maintain a file containing the table of specification, items developed, mark scheme and item classification table, drafts of the test with comments from peers, a printed copy of the test, the result sheet from one class, the compiled result sheet, and two reports.

It is recommended that this assignment account for 20% of the mark for this course.

USE OF STANDARDS IN TESTING

Key words

- Norm-referenced tests
- Criterion-referenced tests
- Standards
- Content-based standard setting
- Performance standards
- SOLO Taxonomy
- Bloom's Taxonomy

Norm- and criterion-referenced tests

Developed by Dr Rizwan Akram Rana

Warm-up questions

Ask the class some warm-up questions to gauge their existing knowledge on norm-referenced and criterion-referenced tests. Questions might include the following:

- What two types of tests are used in educational measurement?
- How would you described norm-referenced and criterion-referenced tests in educational measurement?

Group discussion

Introduce the differences between norms and criteria, and define the terms *norm-referenced* and *criterion-referenced tests*. It may be helpful to provide Student Teachers with some reading material that outlines these concepts, such as pp. 57–61 from Kubiszyn and Borich's *Educational testing and measurement: Classroom application and practice* (2003).

Divide the class into three groups and ask each group to discuss these two types of tests, particularly with regard to interpreting the tests. The groups should categorize different types of tests, such as driving tests, medical examinations, selection tests, or achievement tests, to have a better understanding of the real-world application of these terms. Also, ask the groups to recall the different procedures used to interpret these tests.

Interactive lecture

Conduct an interactive lecture in which you describe the purpose of norm-referenced and criterion-referenced tests (Kubiszyn and Borich, 2003, p. 61) and when these two types of tests should be used to test student performance (Shrock and Coscarelli, 2007, pp. 28–29). Ask questions so that Student Teachers can share their knowledge – and allow you to gauge their understanding of the topic. Consider adding a visual element by using a PowerPoint presentation.

The discussion should cover the following points:

- the purpose of norm-referenced tests
- the purpose of criterion-referenced tests
- comparing norm-referenced and criterion-referenced tests
- when to use norm-referenced tests
- when to use criterion-referenced tests.

Comparing of different scenarios

Have Student Teachers further examine the usability of norm-referenced and criterion-referenced tests.

Present the following three scenarios to the class. Ask the Student Teachers to analyse each scenario and decide whether a norm-referenced or criterion-referenced test should be used. They should explain how they came to their conclusion.

- In 1974, the Government of Pakistan started a national distance education programme by establishing Allama Iqbal Open University in Islamabad. This university used television, audio tapes, and textbooks as learning resources. To pass an exam, a student must attain 33% according to university assessment.
- Students must take an exam in which they demonstrate specific skills in order to become chartered accountants.
- The Higher Education Commission called for papers on educational assessment. The authors of the top five research papers will receive 50 000 Rs./-.

Idea conceived by Shrock and Coscarelli, 2007, p. 29.

Conclusion

Summarize the main points of the lecture and remind the class of the main points covered in the introductory lecture:

- the purpose of norm-referenced tests
- the purpose of criterion-referenced tests
- comparing norm-referenced and criterion-referenced tests
- when to use norm-referenced tests
- when to use criterion-referenced tests.

Homework

Ask Student Teachers to write a brief essay outlining the usability of norm-referenced and criterion-referenced tests in Pakistan.

References

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice* (pp. 57–61). Hoboken, NJ: John Wiley & Sons, Inc.

Shrock, S. A., & Coscarelli, W. C. (2007). *Criterion-referenced test development: Technical and legal guidelines for corporate training* (pp. 28–29). Hoboken, NJ: John Wiley & Sons, Inc.

The contributors would like to thank Tooba Saleem, Shumaila Hameed, Sehar Rashid and Nighat Ara Shams from IER, Punjab University for their contribution to the teaching notes.

Use of standards in education

Developed by Isbah Mustafa

Introduction

Introduce the concept of standards in the context of teaching and, more specifically, the National Professional Standards for Teachers, which were developed in 2009. Engage the class in a discussion on the need for national standards. Responses should address issues such as quality assurance and uniformity across the country.

National Professional Standards for Teachers

Distribute the 'National Professional Standards for Teachers in Pakistan' handout. This document includes the standards developed by the Policy and Planning Wing of Pakistan's Ministry of Education in 2009, which are available in full from:

- http://unesco.org.pk/education/teachereducation/files/National%20Professional%20Standards%20for%20Teachers.pdf

Explain why each standard has a content, disposition, and performance component, and use the standards as an example.

Activity on National Professional Standards for Teachers

Working in groups of three, have Student Teachers appraise the list of 10 standards for their comprehensiveness and relatedness. After the groups have had sufficient time to discuss the standards, they will join with another group to discuss their analysis.

Each group will present the learning outcomes to the class.

Conclusion

Summarize the need for standards in teacher education. Note that in developing a set of standards, effort must be made to ensure that they are both relevant and comprehensive. Explain how each teacher standard is structured; note the heading, the two- to three-sentence description, and learning outcomes that test content, disposition and Student Teacher performance.

Handout

National professional standards for teachers in Pakistan

National professional standards for teachers in Pakistan

Standard 1:	Subject matter knowledge
Standard 2:	Human growth and development
Standard 3:	Knowledge of Islamic ethical values/social life skills
Standard 4:	Instruction planning and strategies
Standard 5:	Assessment
Standard 6:	Learning environment
Standard 7:	Effective communication and proficient use of information communication technologies
Standard 8:	Collaboration and partnerships
Standard 9:	Continuous professional development and code of conduct
Standard 10:	Teaching of English as second/foreign language (ESL/EFL)

Composition of professional standards

Each standard has 3 parts
a. Knowledge and Understandings (Content) What the teacher knows
b. Dispositions Behaviors, attitude and values
c. Performances (Skills) What the teacher can do and should be able to do

From the National Professional Standards for Teachers, p. 9. Available from:

- http://unesco.org.pk/education/teachereducation/files/National%20Professional%20Standards%20for%20Teachers.pdf

Use of standards in testing

Developed by Isbah Mustafa

Standards-referenced test

Explain the difference between norm-referenced and criterion-referenced tests. The former produces results by comparing students with each other and the latter produces result based on a selected criterion (e.g. skill or content area). The skill or criterion serves as the standard on which performance is measured. A criterion-referenced test is considered to be a more precise tool, as it reports student achievement with regard to a specific set of abilities. This also enables it to be used, with different questions on the same topics, to measure change over time or to compare performance within a group. The most common example of a criterion-referenced test is the promotion examination. When a student fails a grade promotion exam, this indicates that the student has failed to achieve the set goals for the grade. More recently, student abilities have been gauged with criterion-referenced tests using standards.

The purpose of standards-referenced testing

Introduce the benefits of standards-referenced tests, such as precision of reporting. Explain how standards-referenced tests show student progress because they relate performance to standards. Standards also facilitate comparison of students in different cohorts, as student performance is matched to standards that are uniform for all. Remind the class of the SLOs used earlier in the test development exercise in which groups developed three sets of items based on the same SLOs. In the same way, different forms of a test can be constructed based on same standards.

Ask Student Teachers to share a standards-referenced test they have come across in their academic career. Using the examples supplied by the Student Teachers, discuss how standards help support testees in achieving the examination's set goal.

Using standards in testing

The National Curriculum 2006 includes competency in content areas, standards, benchmarks, and Student Learning Outcomes (SLOs). Standards are defined in the curriculum as '[defining competency] by specifying broadly, the knowledge, skills and attitudes that students will acquire, should know and be able to do in a particular key learning area during twelve years of schooling' (National Curriculum for English Language Grades I–XII, 2006, p. 5)

Share the 'Standards in testing' handout. Ask Student Teachers to read the table and decide what the test intends to measure. The table clearly indicates that the aim is to maintain assessment focus on the competencies and standards.

Conclusion

End the session by asking the Student Teachers to write a paragraph explaining what standards are and how they are used in testing.

Handout

Standards in testing

Table 6.3: Recommended allocation of marks in relation to standards, benchmarks and SLOs for Grades IX-X

Reference	Benchmarks: language and skill focus (please see detailed benchmarks)	No. of SLOs	Weighting			
			Paper A		Paper B	
			Percentage	Marks	Percentage	Marks
C1, S1	Patterns of text organization	4	15%	11	-	-
	Reading comprehension strategies	7	30%	23	-	-
	Interpretation of visual organizers	5	15%	11	-	-
	Study skills	6	CA	CA	CA	CA
C1, S2	Interaction with literary texts	9	25%	19	-	-
C2, S1	Techniques for effective text organization, development and writer's craft	5	-		15%	11
	Descriptive, narrative, expository, persuasive, analytical writing	8	-		20%	15
	Interpersonal and transactional writing	8	-		20%	15
	Editing and revising	4*	CA	CA	CA	CA
C3, S1	Functions and co-functions*	1	FA	FA	FA	FA
	Group dynamics*	4	FA	FA	FA	FA
C4, S1	Pronunciation development*	4	FA	FA	FA	FA
C4, S2	Vocabulary enhancement	2	15%	11	-	-
C4, S3	Grammatical functions	9	-	-	25%	19
	Mechanics of punctuation	3	-	-	10%	7.5
	Sentence structure	4	-	-	10%	7.5
C5, S1	Attributes for peaceful co-existence	CA	CA	CA	CA	CA
	Valuing self and diversity	CA	CA	CA	CA	CA
	Approach contemporary issues as thinking individuals	CA	CA	CA	CA	CA
Total			**100%**	**75**	**100%**	**75**

CA, classroom activity; C, competency; FA, formative assessments; S, standard. From the National Curriculum for English Language Grades I–XII (2006), p. 159. Available from: http://passpunjabedu.com/index.php?option=com_phocadownload&view=category&id=4:curriculums-2006&Itemid=34

The need for standard setting

Developed by Dr Khalid Saleem

Measuring content and performance

Start the session with a brainstorming activity. Ask Student Teachers how they would judge the performance of a cricket player.

Student Teachers will likely respond that the performance is judged by how well the player has played.

Next ask them how they would know that a student is good in social studies.

Student Teachers will likely respond that a student who has knowledge of social studies is a good student.

Use these responses to show that the differences in measurements applied: performance standards and content standards.

Content and performance standards

Describe various methods used to judge performance in different fields. Explain that an approved way of doing the things is called a performance standard. As noted earlier, a cricketer's performance will be judged on performance standards whereas a social studies student's knowledge will be determined on content standards.

In education, content standards indicate what should be taught and assessed, and performance standards indicate how well the content standards have been met.

Explain content standards by describing how they help to form learning outcomes. As an example, refer to the National Curriculum 2006's SLOs, which have two parts: one that refers to the desired outcome and one that refers to the outcome with regard to the content area.

Next explain performance standards and performance descriptors. (Use the information on the following site from the British Columbia Ministry of Education to inform your lecture:

- http://www.bced.gov.bc.ca/perf_stands/reading.htm

Conclusion

Conclude by asking Student Teachers to list differences between the objectives and structure of content and performance standards.

Setting content-based standards

Developed by Isbah Mustafa

Composing of content standards

Provide Student Teachers with a copy of the 'Selected class 6 SLOs in English, mathematics, and science' handout (from the 'Defining test objectives' section).

Working in pairs, Student Teachers will identify 10 command words and 10 content areas in the SLOs. This activity will help Student Teachers better understand the structure of SLOs.

Once they have identified command words and content areas, ask the pairs to categorize the SLOs based on the level of effort, which includes elements such as the time required for learning, required effort from students, and required effort from teachers. In this case, Student Teachers should focus on teacher effort.

Student Teachers will place the SLOs in three categories: SLOs achievable by teaching one 40-minute class, SLOs achievable by teaching in three or four classes, and SLOs requiring more classes. For example, 'Write a story using the elements of story writing' would fall into the third category, as this requires students to include the following basic elements when writing a story:

- character
- plot
- conflict
- setting
- theme.

It is assumed that without prior knowledge of this area, class 5 students would require more than four classes to be able to achieve this.

Setting content standards

Initiate a discussion on requirements for content standards, particularly the requirements related to selecting targeted content and the involvement of experts. On the board, list these two requirement categories and then write Student Teachers' responses according to the category to which they feel the requirements belong.

Conclusion

Ask for volunteers to summarize the composition of content standards and the essential requirements for setting content standards. Invite other Student Teachers to add to the summaries provided by their peers.

Performance standards

Developed by Isbah Mustafa

Introduction

Explain that performance standards are a further development of content standards. They define expected performance relative to a particular content area or skill. In this context, the term *standard* is used as a set score students need in order to demonstrate that they have the relevant knowledge or skills. Performance standards are defined as '[a]n 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' (from https://www.questionmark.com/us/Pages/glossary.aspx#P).

Performance standards or levels are established along a range of scores. For example, examinees are classified in different levels, such as pass/fail, a range of six marks, or a range of nine marks (stanines).

Using performance standards

Ask Student Teachers to think of the benefits of performance standards and invite them to share their ideas. Engage the entire class in a discussion on these benefits. Make sure that Student Teachers note that performance standards support accountability, as they clearly define what an examinee has achieved in terms of content knowledge or skill.

Performance descriptors

Reintroduce the concept of performance descriptors, which was discussed in the section 'The need for standard setting'. A descriptor consists of a statement of the knowledge, skill, or ability an examinee must display at a particular performance level. It may include sample responses that reflect the level of performance.

The exemplar level of performance for an examinee is determined by considering the borderline examinees. A borderline examinee is one who is barely fits into a given category.

Evaluation of performance descriptors

Share a list of performance descriptors. Examples can be found on the following sites:

- Languages: Assessable elements and descriptors of quality for A–E (2007).
  - http://www.qsa.qld.edu.au/downloads/p_10/qcar_aed_languages.pdf
- Student performance standards (n.d.)
  - http://www.ce.utexas.edu/prof/mckinney/StudentPerformanceStandards.htm
- GPS by grade level, K–8 (2011).
  - https://www.georgiastandards.org/standards/Pages/BrowseStandardsGPS_by_Grade_Level_K-8.aspx

Working in groups, Student Teachers will judge the descriptors' suitability for primary students in Pakistan. Each group will present their evaluation in class.

Alternatively, Student Teachers can ask the English teacher at their school for pieces of student writing that meet the class 5 performance standard, 'Write a story using the elements of story writing'. Student Teachers will review these pieces and decide how the performance standard could be improved so that an agreement on standards is reached.

Cut score

Explain that a critical step in the development and use of performance standards on tests is to establish one or more cut points. These cut points divide the score range into performance levels, or grades scores. Cut scores are selected points on the score scale of a test. They represent the rules by which tests are interpreted.

Methods of performance standard setting

Developed by Isbah Mustafa

Performance standard setting

Discuss the procedure for performance standard setting. Hambleton and Ritanick (2000) list the following steps in the procedure:

1) Select a standard-setting method.
2) Choose a panel and design.
3) Prepare descriptions of performance categories.
4) Train panellists to use the method.
5) Collect item ratings.
6) Provide feedback and facilitate discussion.
7) Compile panellist ratings and obtain performance standards.
8) Conduct panel evaluation.
9) Compile validity evidences and prepare technical documentation.

From Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. In R. L. Brennan, *Educational measurement* (4th ed.). Westport, CT: National Council on Measurement in Education and American Council on Education.

The most common standard-setting methods

Introduce the two most common ways to set standards:

- test-centred standards
- examinee-centred standards.

Ask Student Teachers to think about how these two methods differ from each other. Invite them to share their thoughts and then elaborate on them by filling in any gaps.

Test-centred standard setting

Introduce the Angoff Method, a method frequently used to set standards. This method is used for multiple-choice items. A panel examines each question individually and determines the probability of a borderline student answering the question correctly. The number of probable correct responses from a borderline student is then totalled. The responses from all members of the panel are averaged to determine the performance standard. The panel then discusses its determinations to come to a consensus.

Many modified versions of the Angoff Method have used since the method was introduced in 1970s.

Consider introducing other test-centred methods, such as Ebel's approach.

Examinee-centred standard setting

Introduce examinee-centred standard setting, which requires rating a sample of candidates. The two most common methods are the borderline group method and the contrasting groups method.

Borderline group method

This is the conventional practice followed by senior examiners setting O-level grade boundaries. Not every borderline is based on human judgement because it would be too time-consuming. Usually A/B, C/D, and E/F are identified in this way, but B/C and D/ E are consider the midpoints in the A/B to C/D ranges.

Contrasting groups method

The contrasting groups method calls for scrutiny of the performance of groups identified as 'masters' and 'non-masters' based on evidence external to the test itself. It is difficult to simulate without actual data. The closest Student Teachers could get would be to administer their test at its intended grade level, the grade below (for non-masters), and the grade above (for masters).

Ask Student Teachers to think about how performance standards could be developed using these two methods. Use the discussion to give additional details about these two methods.

Conclusion

Ask for volunteers to summarize the Agnoff Method procedure and examinee-centred methods of standard setting.

Practise setting performance standards

Developed by Isbah Mustafa

In this activity, Student Teachers will perform a standards-setting exercise to understand the procedure and, thus, its potential use in Pakistan as content standards in primary, middle, and secondary school.

Prerequisites

For the practicum, Student Teachers will need the test papers, answer booklets, compiled results, and class results prepared during the 'Test development' section.

Selecting a method to set performance standards

Ask Student Teachers to work in the same group of six as they did in the 'Test development' section.

For the practicum in this unit, two groups will work together. Within the larger groups, one group will work on examinee-centred performance standard setting and one will work on test-centred performance standard setting. Each group will determine the material and information required to apply the method (e.g. multiple-

choice items for the Angoff Method, marked scripts for the examinee-centred method). Each group will write the procedure they think should be used to apply the method.

For homework, each group of six will do additional research online and in the library to confirm that they have outlined the steps correctly. They should also ensure that they have the necessary materials to conduct the exercise in the next session.

Based on each member's research, the group of six will finalize their list of requirements. The groups should also assign tasks to various members (e.g. someone to do the calculations, a record keeper).

The groups will calculate scores in order to determine the standards. Each group will first judge the identified standard and then share it with the class.

As homework, each group will prepare a report on the procedures involved in performance standard setting and what is appropriate for primary school classes. Group members will need to discuss their work and reach a consensus on their recommendations. For each method, the report should include a description, the requirements, and the necessary procedure. Also, explanations should be provided that outline why each method is appropriate in a given context.

Handout

Standardized tests

Handout developed by Dr Thomas Christie

Teachers have access to just one or two classes in a particular school, and after a while, they may come to expect and accept their students' performance as normal. In some cases, teachers may lose perspective. This is why external bodies must perform moderation exercises before they accept teacher assessments. The external body has access to results from many schools and is in a position to recognize huge differences in average performance between classes, even if teachers think their own class is 'normal'.

Standardized tests overcome this difficulty. They are piloted with carefully selected samples from a national or regional population of students. Many schools and classrooms are represented. The distribution of scores obtained by this sample of students (usually several thousand) is then scaled to a predetermined score range. For example, the observed average may be adjusted to 50, with a standard deviation of 10, so that just a few students score more than 80 or less than 20. These scores can then be interpreted in terms of examinee-centred standards: of 100 examinees, half will score 50 and only three will score above 70. Standardized tests, which usually have to be purchased to cover the costs of standardization, help classroom teachers understand their students' performance in national terms. There are international standardized tests, such as Trends in International Mathematics and Science Study (TIMMS), that allow a nation to see how it stands in international terms. Student Teachers may wish to discuss why Pakistan does not participate in these international comparisons.

Standardized tests

Review the contents of the handout and discuss the following questions:

- What is a standardized test?
- What is the process of test standardization?
- Why are there no Pakistani standardized tests?

Taxonomy in testing: Using Bloom's Taxonomy in test development

Divide the Student Teachers into groups of five and ask each group to recall the concepts of higher- and lower-order thinking. The groups should brainstorm and discuss the level of student understanding necessary for achievement or mastery at the end of primary and secondary school.

Bring the class back together, and give a member from each group one minute to share their ideas with the entire class.

In a mini-lecture, connect Student Teachers' ideas about the levels of student understanding with Bloom's Taxonomy. Explain how Student Teachers can choose appropriate Bloom's levels for test items.

Handout

Recognizing Bloom's Taxonomy level

List A is a set of items from the AKU-EB Secondary School Certificate Part 1 examinations.
List B provides the SLOs that the items test. Match each item with the correct SLO.

Grade IX student learning outcomes and test items

List A: Test Items	List B: SLOs of Knowledge (K), Understanding (U) and Application (A) level
Which phases of the cell cycle are collectively called interphase? A. G1, G2, and M B. S, G2, and M C. G1, G2, and S D. G1, M, and S	Identify food sources and metabolic functions of vitamins A, C, and D (K)
An invertebrate is found with the following characteristics: • It lives only in sea water. • Its body is covered with spines. • Its internal skeleton is present. In which phylum should it be placed? A. Mollusca B. Nematoda C. Coelenterata D. Echinodermata	Describe the salient features of invertebrates using paramaecium, sycon, jellyfish, tapeworm, roundworm, earthworm, snail, butterfly, and sea star as examples (U)
In pea plants the flowers can either be purple or white. The allele for purple is P whereas the allele for white is p. If a heterozygous purple plant is crossed with a white flowered plant, what percentage of purple and white flowered plants do you expect in the F1 generation? A. 100% purple B. 75 % purple, 25 % white C. 50 % purple, 50% white D. 100 % white	Explain the relationship between biology and other branches of science (physics, chemistry, mathematics, geography, and economics) and relate the relationships with suitable examples from daily life (U)
A man has blood group AB and his wife has blood group O. What are the possible blood groups of their children? A. AB and O only B. A and B only C. A, B, and O only D. A, B, and AB only	Describe complete dominance using the terms dominant, recessive, phenotype, genotype, homozygous, heterozygous, P1, F1, F2 generations and proving it diagrammatically through a monohybrid genetic cross (A)

Grade IX student learning outcomes and test items (cont.)

List A: Test Items	List B: SLOs of Knowledge (K), Understanding (U) and Application (A) level
What should a person suffering with bleeding gums and poor healing of wounds include in his/her diet? A. Dairy products B. Citrus fruits C. Red meat D. Cereals	Describe complete dominance using the terms dominant, recessive, phenotype, genotype, homozygous, heterozygous, P1, F1, F2 generations and proving it diagrammatically through a monohybrid genetic cross (A)
The conversion of insoluble fibrinogen into fibrin threads is carried out by enzyme thrombin, which works in the presence of vitamin K only. Here, vitamin K is acting as a(n) A. substrate B. active site C. co-factor D. inhibitor	Explain why some enzymes require co-factor for their functioning (U)
The production of insulin on a large scale through bacteria has made insulin inexpensive and easily available to diabetic patients. This indicates the relationship of biology with A. mathematics B. economics C. geography D. physics	Define cell cycle and its main phases: interphase and division (K)

Taken from examination syllabus and examination papers of Aga Khan University Examination Board.

After the class completes this handout, discuss the answers and engage them in a discussion about how they matched the items and SLOs.

SOLO Taxonomy and test development

Developed by Dr Muhammad Sarwar

What is the SOLO Taxonomy?

In a mini-lecture, explain the concept of Structure of Observed Learning Outcome (SOLO) Taxonomy and its five levels of understanding for a given topic or area of learning.

Consider using the article 'SOLO Taxonomy: Giving students a sense of progress in learning', available from: http://edu.blogs.com/edublogs/2012/08/solo-taxonomy-giving-students-a-sense-of-progress-in-learning.html, as a resource in preparing the lecture.

Divide Student Teachers into group of four or five, and ask them to brainstorm the use of SOLO Taxonomy in evaluating student learning relative to Bloom's Taxonomy.

After the class has finished brainstorming, engage the class in a whole-class discussion about the advantages of the SOLO model for evaluating student learning.

Determining the behaviours to be assessed

Developed by Dr Muhammad Sarwar

Introduction

Ask Student Teachers what a learning outcome is. Write the salient points they make on the board.

Explain that SLOs are statements that specify what students should know, be able to do, or be able to demonstrate when they have completed a programme, activity, course, or project. Outcomes are usually expressed as knowledge, skills, attitudes, or values. SLOs use a command word to specify student actions that must be observable and measurable demonstrations of an outcome.

Interactive lecture

Introduce three types of learning outcomes:

- cognitive learning outcomes (knowledge or mental skills)
- affective learning outcomes (growth in feelings or emotional areas)
- psychomotor learning outcomes (manual or physical skills).

CHARACTERISTICS OF A GOOD TEST

Key words

- Item analysis
- Test reliability
- Test validity
- Test utility

The concept of a good test

Divide Student Teachers into groups of four or five. Ask each group to list the key features that they believe are essential to a good test. Their responses should be based on their experiences as students. Each group will then share their list in the class.

Summarize the common features of a good test as discussed by the Student Teachers, and then elaborate on the concept of a good test through a mini-lecture, literature review, and interactive group discussion. At the end of the session, ask Student Teachers whether they can objectively judge if a test is good or bad.

Item analysis

Developed by Samreen Jalal

Warm-up questions

Review material from previous sessions by asking Student Teachers the following questions:

- How is the difficulty level of a test item determined?
- How is the discrimination index of a test item determined?
- How do you identify the plausibility of a test item?

Group discussion

Briefly describe the three elements of item analysis —item difficulty, discrimination, and plausibility—and provide Student Teachers with materials on the test development process and the use of item analysis in classroom assessment. (To prepare materials for class, use Ebel, 1991, pp. 220–233, and Nitko, 1996, pp. 308–317.)

Divide Student Teachers into four groups. The groups will discuss the materials to gain a better understanding of the concepts.

Present one of the topics listed below to each group:

- computation of an item difficulty index
- computation of an item discrimination index
- item selection
- item revision.

Ask each group to present their findings on the topic to the rest of the class. During the presentations, act as a facilitator and provide assistance when Student Teachers have difficulty explaining their points.

As an extension of this activity, develop and present an interactive lecture based on the points made by Nitko (1996, pp. 312–317). Consider using a visual presentation (e.g. PowerPoint presentation) to supplement the lecture. Invite Student Teachers to respond to questions and ask their own. Also, make sure to discuss the following during the lecture:

- using item difficulty information to modify items
- using the discrimination index to improve test items
- using guidelines to select items for classroom tests.

Homework

Student Teachers will perform an item analysis using the question level student scores obtained in the 'Test development' unit.

Reference

Nitko, A. J. (1996). *Educational assessment of students*. Upper Saddle River, NJ: Prentice Hall.

Test reliability

Developed by Dr Muhammad Sarwar, Dr Shafqat Hussain, and Dr Bashir Hussain

Introduction

Explain the fundamental concepts of reliability and validity through a demonstration with a target. Draw concentric circles on the board, with a target in the centre. Provide a piece of chalk (or something that will leave a temporary mark) to three volunteers, and ask them to hit the target from a fixed distance. Note how close each came to the target, and then explain whether they were consistent (i.e. reliable) and/or accurate (i.e. valid) in hitting the target.

Introduce the concept of reliability as consistent outcomes on repeated occasions. Explain the various types and methods of reliability analysis and discuss strategies for improving test reliability. Tell Student Teachers to be prepared to calculate reliability in the next session.

Practice session to calculate reliability of tests

Initiate a discussion by asking the Student Teachers to list possible advantages of having alternate or parallel forms of the same test, particularly from the perspective of an examinee. Ensure that Student Teachers make points related to reliability and that the following terms are introduced:

- interrater
- split-half, internal consistency
- test-retest.

Next explain ways to calculate reliability. Provide some test score data to Student Teachers and ask them to review the information in the context of reliability (i.e. calculate various forms of reliability). As Student Teachers work with the figures, walk through the room to help them with the task. This will help you assess whether Student Teachers are able to list and describe various measures of reliability and to identify the strengths and weakness of each.

Student Teachers should note that all forms involve a correlation between repeated attempts or repeated judgements: each one seeks consistency across repeated attempts.

Test validity

Developed by Dr Rizwan Akram Rana

Brainstorming

Ask Student Teachers the following questions on test validity:

- What are the characteristics of a good test?
- How is test validity defined?
- Why is a test required to be valid?
- What methods ensure the validity of a test?

Introduction

Introduce the topic of test validity with a brief description. Ensure that Student Teachers understand the difference between validity and reliability.

Interactive lecture

Distribute a handout on test validity and related concepts. (Either use the resources listed at the end of this section or create a handout based on them.) Give Student Teachers 10 minutes to review the materials.

After this, explain the concept of test validity and its importance in assessment and test development. Also, explain content, construct, concurrent and predictive validity, and methods to ensure test validity. Provide example of factors that affect test validity.

Group work

Divide Student Teachers into groups of four or five. The groups should discuss the material covered in the interactive lecture. Each group should compile a brief list of factors that can affect test validity, which they will then present to the entire class. The rest of the class and Instructor should provide feedback on whether the factor is likely to affect validity.

Conclusion

Summarize the main points of the lecture as well as the points made by the Student Teachers.

Assessment

Provide an example of a test you have developed. Ask the Student Teachers to establish its content validity and to write about the steps they would take to ensure the validity of a classroom test.

References

Abell, L., Springer, W. D., & Kamata, A. (2009). *Developing and validating rapid assessment instruments* (pp. 98–103). New York: Oxford University Press.

Miller, M. D., Linn, R. L,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.) (pp. 70–106). Chicago, IL: R. R. Donnelley & Sons Co.

Phelan, C., & Wren, J. (2005–06). Exploring reliability in academic assessment. Retrieved from:
➢ http://www.uni.edu/chfasoa/reliabilityandvalidity.htm

The contributors would like to thank Tooba Saleem, Shumaila Hameed, Sehar Rashid, and Nighat Ara Shams from IER, Punjab University for their contribution to the teaching notes.

Relationship between test reliability and test validity

Developed by Dr Rizwan Akram Rana

Warm-up activity

Ask Student Teachers the following questions:

- What can you conclude from the consistency of scores?
- Why is consistency sometimes required? Why is it sometimes undesirable, especially in achievement tests?
- What procedures are used to ensure the reliability of students' scores?
- What are some factors that affect the consistency of students' scores?

Introduction

Introduce the notion of the relationship between test reliability and validity by providing definitions of each term. Make sure that students are clear about the distinction between these concepts.

Interactive lecture

Distribute a handout on test reliability and its importance in developing exams. (Either use the resources listed at the end of this section or create a handout based on them.)

After Student Teachers have had time to review the handout, explain the different types of test reliability. The main point should include:

- test reliability
- the relationship between reliability and validity
- the different types of reliability
- methods for establishing external and internal consistency and reliability
- the interpretation of the reliability coefficient.

Group work

Divide the class into groups of four or five, and ask them to discuss the points made in the interactive lecture. Each group will be required to calculate the internal consistency reliability of a classroom test, based on data provided by the Instructor.

Conclusion

Summarize the main points of the lecture as well as the points made by the Student Teachers.

Assessment

Ask Student Teachers to review test reliability at home and write 20 multiple-choice items on any subject of their choosing. Student Teachers should administer the test in an appropriate classroom by applying relevant procedures to calculate the test-retest and the split-half, internal consistency.

References

Abell, L., Springer, W. D., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. New York: Oxford University Press.

Miller, M. D., Linn, R. L,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.). Chicago, IL: R. R. Donnelley & Sons Co.

Phelan, C., & Wren, J. (2005–06). Exploring reliability in academic assessment. Retrieved from:

- http://www.uni.edu/chfasoa/reliabilityandvalidity.htm

The contributors would like to thank Tooba Saleem, Shumaila Hameed, Sehar Rashid and Nighat Ara Shams from IER, Punjab University for their contribution to the teaching notes.

Types of reliability

Developed by Dr Rizwan Akram Rana

Brainstorming

Review aspects of reliability by asking Student Teachers the following questions:

- What is test reliability?
- Can you differentiate between the various types of test reliability?
- What are some factors that affect the consistency of students' scores?

Introduction

Introduce the concept of rater (scorer) reliability as well factors affecting test reliability. Briefly describe different types of internal consistency reliability procedures, reader and rater (scorer) reliability, and factors affecting the reliability of test scores.

Reading and discussion

Distribute a handout on rater (scorer) reliability and factors that affect score reliability, including rubrics. (Either use the resources listed at the end of this section or create a handout based on them.) Use the handout as the basis for discussion.

Group work

Divide Student Teachers in groups of four or five, and provide the groups with data about test results. Each group will develop a scoring rubric for a common question and then score the same six pupils' responses to the question. Each group will be required to calculate the rater (scorer) reliability within the group and the reliability between groups using group mean scores.

Conclusion

Summarize the main points of the lecture.

Assessment

Ask the Student Teachers to study different concepts of test reliability at home. They should then write five essay items with scoring rubrics on any subject of their choosing. Student Teachers should administer the test in an appropriate classroom and apply relevant procedures to calculate scorer reliability.

References

Abell, L., Springer, W. D., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. New York: Oxford University Press.

Gardner, J. R. (Ed.).(2005). *Assessment and learning*. London: Sage Publisher.

Miller, M. D., Linn, R. L,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.). Chicago, IL: R. R. Donnelley & Sons Co.

Phelan, C., & Wren, J. (2005–06). Exploring reliability in academic assessment. Retrieved from:

- http://www.uni.edu/chfasoa/reliabilityandvalidity.htm

The contributors would like to thank Tooba Saleem, Shumaila Hameed, Sehar Rashid and Nighat Ara Shams from IER, Punjab University for their contribution to the teaching notes.

Utility of test

Developed by Dr Muhammad Sarwar, Dr Shafqat Hussain, and Dr Bashir Hussain

This session will focus on the application of validity, reliability, objectivity, adequacy, and differentiability. It aims to ensure that Student Teachers are able to correctly score, interpret, and provide feedback on an individual's test results. This session should also cover reporting tests (marks vs grades) and introduce the concept of the standard error of measurement.

Test reliability and validity and standard error of measurement

Student Teachers will practise and apply their understanding of test reliability and validity as it relates to the standard error of measurement. In the practice session, Student Teachers should demonstrate that they can integrate test results with other information about the testee, such as other test and assessment results or performance/behaviour in different educational settings.

Providing feedback

Instruct Student Teachers on how to provide appropriate feedback regarding test results, both verbally and in writing, to the testee and other stakeholders, such as parents, teachers, or other professionals.

Conclusion

Conclude the session with the summary of the characteristics of a good test.

References

Cohen, R., & Swerdlik, M. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7th ed.). McGraw-Hill.

FitzGerald, J. (1999). Testing the tests. Retrieved from:
➢ http://www.actualanalysis.com/tests.htm

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Miller, M. D., Linn, R. L,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.). Chicago, IL: R. R. Donnelley & Sons Co.

Reliability & validity. Retrieved from:
➢ http://www.socialresearchmethods.net/kb/relandval.php

GRADING AND REPORTING

Key words

- Statistics
- Modes of interpreting data
- Measures of central tendency
- Measures of variability
- Functions and uses of grades
- Types of grading
- Principles for managing the grading process
- Interpretation of test scores
- Grading criteria
- Grading standards (description of grades, checklist, rubrics)
- Modes of reporting
- Problems in grading and reporting
- Item analysis

The importance of statistics and modes of interpreting data

Developed by Samreen Jalal

Introduction

Creating statistics involves techniques and procedures to summarize information that can be stated numerically. There are two broad categories: non-parametric statistics, which deal with counts of observations that make assumptions about probability, and parametric statistics, which assume that observations can be measured on a scale. Scales for parametric statistics have units that are observable, like yards or metres, or attributes made observable by measurement, such as ability or achievement.

Explain that knowledge of statistics will help Student Teachers to collect, organize, and analyse data in order to make decisions about the teaching-learning process.

Statistics have many benefits, including the following:

- If statistics are examined and analysed properly, then a teacher will have the power to improve in the areas where students appear the weakest.
- Statistics enable teachers to monitor student progress throughout the school year.
- Statistics help researchers analyse the results of their studies.
- Statistics are key to the behavioural sciences, which are based on empirical research.

Q and A session

Ask Student Teachers the following questions in order to check their understanding of statistical concepts:

- How many types of statistics are there?
- Which type of statistics deals with simple numbers of observations?
- Which type of statistics is used by teachers to summarize results on report cards?
- How we can interpret and analyse educational data?

Measures of central tendency

These are simply ways of identifying the most typical observation in a distribution.

The mode is the one that occurs most often. It can come anywhere in the distribution. In a very difficult test, it might be a score of zero. In a very easy test, it might be 10 out of 10. In a good test, it will be somewhere near the middle.

The median is the middle score and is usually reported in non-parametric statistics where each unit is indivisible.

The mean is parametric. However, its application can be difficult at first glance. For example, the average British family has a mother, a father, and 2.5 children, making

for a family of 4.5 people. While this does not worry a parametric statistician, it may worry you. What does half (0.5) of a child look like? The non-parametric version would have two adults and two children as the mean family size. Note that this is still true today, even though 30% of children come from single-parent homes. Non-parametric statistics are not good at registering change.

Measures of variability

Developed by Samreen Jalal

Q and A session

Assess Student Teachers' knowledge of variability measures by asking the following questions:

- Which type of variability is found by subtracting the lowest number from the highest number?
- How do you find the semi-interquartile range based on the median of the top and bottom halves of the data?

Instructor presentation

Introduce measures of variability and provide examples of each. A measure of variation describes how spread out a data set is or the distribution of data. It is also known as a measure of dispersion or a measure of spread. A measure of variation indicates how 'spread out' the data set is. In non-parametric statistics, the measures include the median range and the interquartile range. The key measure of variation is the variance, which is the standard deviation squared.

Explain and calculate these measures by giving the following examples:

- **Range** is the simplest measure of dispersion. It is defined as the difference between the largest value and the smallest value in the data set: Range $= X_{max} - X_{min} = 109.5 - 9.5 = 100$.
- **Quartile** deviation, or semi-interquartile range, is defined as half of the difference between the third and the first quartiles: $QD = Q_3 - Q_1$.

Presenting data

Introduce different ways to represent observations in tables or graphs. The focus of the presentation should be the frequency of outcomes, which is represented on a frequency distribution. Most natural phenomena, such as the number of flowers on a bush, the numbers of bees in a hive, or the heights of children in a class, are distributed in the same way.

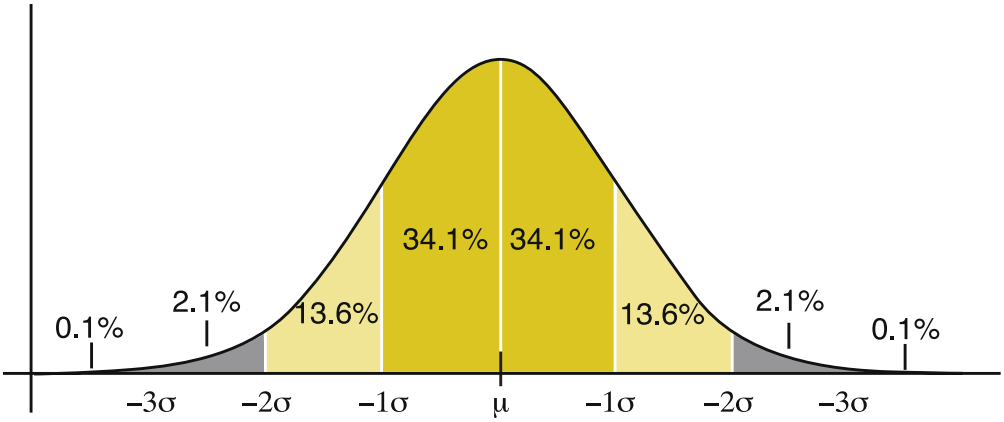
Most observations will cluster around a central value, with progressively fewer cases at the more extreme values. This distribution is so common that it is called the *normal curve*. The normal curve was created based on a formula identified by a 16th-century mathematician named Gauss. (The normal curve is also referred to as the Gaussian curve.) The most useful feature of the normal curve is that the number of observations expected in each part of the curve is generally known. This is calculated in the table of areas under the normal curve, which are reported to four decimal places.

The three measures of central tendency—mode, median, and mean—coincide with the middle of the normal curve. Between the point of inflexion, where the curve changes direction, and the central peak (or the mode), 32% of observations appear. The distance between these points on the scale is the standard deviation (SD). The mean and SD are the parameters in parametric statistics. We use the SD to convert scores on achievement to the chances of them occurring. Once it has been converted on the scale, and thus standardized, it becomes a constant like metres or kilograms, and we locate where we are on the SD curve in units on the baseline. Two SD from the mean along the baseline covers another 16% of observations, leaving just over 2% in the tail. Extending the scale to three SD covers 49.8% of observations. The normal curve never reaches its baseline. There is always the possibility of a more extreme observation that has yet to be seen.

The normal curve belongs to the world of perfect forms. It is a model that indicates the meaning of a mark, providing that the mean and SD of the marks are known. Once we know the mean and SD, the normal curve allows us to make an interpretation of a score.

For example, intelligence quotient (IQ) traditionally has had a mean of 100 and a SD of 15. (This stems from the original Binet–Simon test, which had a SD between 13 and 16, depending upon the children's age.) This tells us that scores of 145 (+3 SD) are very rare—they are so rare that most tests have 140+ as the highest measured IQ.

Discuss the normal curve with Student Teachers



Ask the Student Teachers to consider the following example:

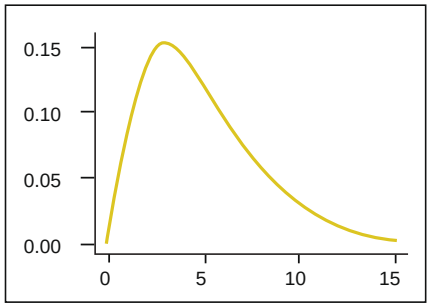
A survey about dog ownership was taken in 20 homes. In each of the 20 homes, people were asked how many dogs they had. The results were recorded as follows: 1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0.

These data can be presented in a cumulative frequency distribution table.

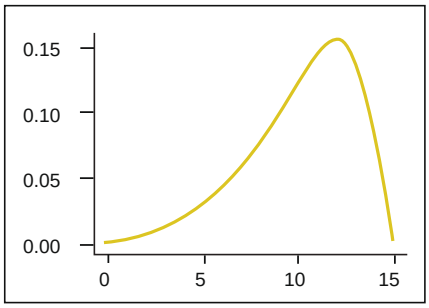
Number of dogs	Frequency (number of observations)	Cumulative frequency	Cumulative proportion
0	4	4	0.20
1	6	10	0.50
2	5	15	0.75
3	3	18	0.90
4	2	20	1.00
5	0	20	1.00

In real life, the data are skewed. For most people, one dog is enough. If the distribution of dogs is to be related to the normal curve, we simply need the number of observations for each number of dogs. Try to draw it as a histogram.

Which of the following curves does the histogram resemble?



Positive skew



Negative skew

Conclusion

Briefly summarize the points discussed.

Reference

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice* (7th ed.) (pp. 235–250). Hoboken, NJ: John Wiley & Sons.

Handout

Measures of central tendency

The following are the final averages of 25 students in a maths class:

46	64	72	79	89
49	66	74	79	91
53	66	75	80	94
60	67	76	83	95
61	71	79	88	98

1) What is the mode?
2) What is the median?
3) Calculate the lower quartile, Q_1, and the upper quartile, Q_3.
4) Calculate the semi-interquartile range.

Answer the following questions based on this data set: 5, 7, 11, 12, 13, 18.

5) What is the mean?
6) Add 8 to each data item. What is the new mean?
7) Subtract 6 from each data item. What is the new mean?
8) Multiply each data item by 7. What is the new mean?
9) Divide each data item by 5. What is the new mean?

Measures of central tendency

Developed by Samreen Jalal

Answers: Measures of central tendency

Provide the answers to the handout so that Student Teachers may assess their performance.

1) 79
2) 75
3) 0.5(25) = 12.5 → round to next higher integer 13 → 13th number is median = 75
4) $Q_1 = 64$, $Q_3 = 88$, semi-interquartile range = 12
5) 10
6) 19
7) 5
8) 77
9) 2.2

Ask Student Teachers what conclusions can be drawn about the mean.

References

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice* (7th ed.) (pp. 250–259, 263–272). Hoboken, NJ: John Wiley & Sons.

Miller, M. D., Linn, R. L,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.) (pp. 501–505). Chicago, IL: R. R. Donnelley & Sons Co.

Grading: Uses, functions, and types

Developed by Samreen Jalal

Q and A session

Ask Student Teachers the following questions:

- How can we display students' progress, strengths, and weaknesses? (Refresh Student Teachers' memories about item-based and examinee-based standard setting.)
- What should the distribution of grades be? How should it be determined?

After a brief discussion, ideas on grading may be analysed and organized by the Instructor in light of how useful they might be in reporting student performance.

Presentation

Introduce different types of grading: item-based (criterion-referenced) using a pass/fail grading and examinee-based (norm-referenced) using actual scores transformed by the normal curve. Provide illustrations of each.

Explain the functions, uses, and types of grading.

Consider using visual aids for the presentation, such as charts or a PowerPoint presentation.

Conclude by summarizing the concepts discussed.

References

Dalumpines, L. (2011). Evaluating learning outcomes: Grading and reporting students progress. Retrieved from:

- http://education-teaching-careers.knoji.com/evaluating-learning-outcomes-grading-and-reporting-students-progress

Grading systems. (n.d.) Retrieved from:

- http://www1.umn.edu/ohr/teachlearn/resources/grading/index.html

Miller, M. D., Linn, R. L., & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.) (pp. 368–371). Chicago, IL: R. R. Donnelley & Sons Co.

Oosterhof, A. (2003). *Developing and using classroom assessments* (3rd ed.) (pp. 216–218). Upper Saddle River, NJ: Merrill Prentice Hall. Available at:

- http://cfmebooksrof.jimdo.com/2013/07/10/developing-and-using-classroom-assessments-3rd-edition-ebook/

Principles and strategies for managing the grading process

Developed by Dr Thomas Christie

The SSC examination in Pakistan reports has a total of 1100 marks based on eight subjects in defined subject groups: premedical, pre-engineering, humanities, and commerce. O-level examinations report grades in individual subjects on a seven-point scale: A*, A, B, C, D, E and Fail. What are the advantages and disadvantages of these two reporting methods?

The SSC reporting method leads to false distinctions. If candidates' exams were marked by a different marker from the same group of markers, it is likely that very few candidates would get the same mark. A candidate's score may vary by 40 or 50 marks if evaluated by a different marker. An award of 990 the first time could be considered worthy of 950 or 1040 by a different marker. As such, it is risky to assume that someone who scored 801 is more capable than someone who scored 799.

If the SSC were to introduce score grades, say 100 marks in each grade, then most of this confusion would disappear, but there would be potential difficulties for those with borderline scores.

Score	Grade
900–1100	A1
800–899	A
700–799	B
600–699	C
500–599	D
400–499	E
0–399	F (fail)

For example, the candidate with 801 marks would receive an A and would be convinced he was a better candidate than the one with 799 and a B. Most other people would agree even though there was no real difference.

O-level grades are designed to avoid this problem by having senior examiners inspect the performance of all candidates who fall one or two marks below an important borderline, usually A*/A, B/C, and E/F, and promote those who seem to belong with the group above.

SSC uses item-based criteria. O-levels use examinee-based criteria. The article in the following handout examines the matter further. The handout should be given to Student Teachers as home reading.

Handout

Variation in Standards in Pakistani School Examination Boards

This paper was presented by Dr Thomas Christie in 2012 at the Aga Khan University Institute for Educational Development Pakistan Conference, In search of relevance and sustainability of educational change (1–3 November 2012).

Variation in standards in Pakistani school examination boards

T. Christie, Director, Aga Khan University Examination Board (AKU-EB)

The research opportunity

Some 4000 students apply for entry to AKU's MBBS degree every year although there are only 100 places available. These students are unlikely to consider themselves average. They are applying for the most sought after course in one of the three top ranked universities in Pakistan. Moreover, entry is 'needs blind' so financial considerations need not influence the decision to compete except in relation to the not inconsiderable cost of sitting the Aga Khan University Admissions Test (UAT). Student motivation can be taken to be fairly uniform. The admission process is in two stages. In Stage-1 the top scoring 10 percent of applicants in UAT are identified, 431 in 2011, 320 in 2012. At this stage the selection decision is based purely on test score: no demographic data are available to the constructors and scorers of UAT. The selected candidates then go into a detailed second phase scrutiny which comprises school performance, achievement in external examinations and two independent interviews to identify the 100 candidates who will enter the University. This research explores how the secondary school examination system conditions candidates' chances of accessing a highly desired educational opportunity. Final admission decisions will have taken account of Higher Secondary School Certificate (HSSC) and A level examinations but here we are simply concerned with how preparation for different examinations at the end of secondary school stage, Grade X for SSC and grade XI for O level, relates to performance in the UAT.

The University Admissions Test (UAT)

Table-1 seeks to establish the credentials of the University Admission Test as a measure of educational potential. There are two English essays, one double marked, to yield a total possible score of 18 derived from a common six band scale ranging from Novice with ineffective communication of ideas to Superior with fluent and natural use of English with minimal errors. The English MCQ tap reading skills. The science items derive from the syllabus elements common to the 2006 revision of the National Curriculum for grades XI and XII and the Cambridge A level topics and outcomes. This section alone deploys negative marking.

The third group, the reasoning items, are of the type commonly referred to as 'word problems' in Mathematics examinations. In both 2011 and 2012 science reasoning items have the highest correlation with the total UAT score. Not that the superiority is significant but the repetition in two different tests of the top place is a very strong

indication that the UAT calls for more than the mere recall of information. Table-2 displays the first principal component of these two incompletely parallel versions of the UAT. A principal components analysis indicates how many underlying attributes are being measured but not what they are. Between 2011 and 2012 there was one change of UAT structure. While in the short writing task, "Explain the meaning of the given Urdu saying in a paragraph", only the saying differed, in the argumentative essay there was a change of format as well as topic. Both topics were well rehearsed in newspaper op-ed pieces, 2011 from economics and 2012 from medical ethics, but the economics argument offered no ideational support whereas in 2012 candidates were given a range of eight assertions to draw upon. The essays were more homogeneous as a result and the writing loading in Table-2 reduced in consequence. The fluctuation in physics is also a function of reduced variance. The physics test in 2011 was too demanding and therefore contributed less variance to the UAT as a whole. Nevertheless although the analysis is sensitive to these variations the test does remain stable overall. Clearly one can entertain a range of hypotheses about what is being measured but the test constructors set out to capture readiness for medical education and it seems not unreasonable to suppose that the UAT is measuring the ability to think through problems encountered in short order, some of them for the very first time, just as in an Accident and Emergency Unit. However the problems set have to be sufficiently challenging to identify the top ten percent of a highly (self) selected population. These candidates are part of the intellectual capital of the nation. The creation of intellectual capital is traditionally the guiding purpose of schools and the recognition of these outcomes is the task of school examination boards working within the confines of the social expectations codified in the national curriculum.

Comparison of academic standards

There are three secondary school curricula in circulation in Pakistan, the National Curriculum of 2002 and earlier which was characteristic of the SSC while the candidate group under review was being educated, the revisions of 2006 through 2008 which are now recognised in part through revised question paper formats but implemented in full only by the two boards with a national remit, Federal Board of Intermediate and Secondary Education (FBISE) and Aga Khan University Examination Board (AKU-EB), and finally the GCE syllabuses examined by Cambridge International Examinations (CIE). Admission to a university medical school is conditional in all three curricula on a common subject combination, the premedical group, which comprises English, Urdu, Mathematics, Pakistan Studies, Islamiyat, Physics, Chemistry and Biology.

In the Boards of Intermediate and Secondary Education (BISE) which run the national school achievement certificates, great reliance is placed on the syllabus, or to be precise, the single text book approved for the course. Any departure there from can render the examination null and void. The FBISE, AKU-EB and GCE examine syllabuses and do not disallow any textbook. The choice of learning material is at the discretion of the schools. The BISE submit to further control of examination format with agreed numbers of MCQ, short response, essay and where appropriate practical marks with rather looser control of question choice while in O levels question choice varies from subject to subject from none at all in O level mathematics to a free for all in the general paper. Within these restrictions the marks of the BISE are treated as

fixed and unchallengeable, grades are derived mechanically from the marks and any suggestion that they might be scaled is seen as tampering with a revealed truth. The grade boundaries, 80% of available marks down to 33% are treated as benchmarks, a kind of gold standard "untainted by values, culture or power" (Bloxham & Boyd, 2012, p.617). Certainly the candidates are essentially powerless. They cannot appeal against the marks assigned by the BISE.

There is only one softening of this absolutist stance. The award of an SSC is based on the raw mark total of eight subjects though failure in any one means failure overall. However it is well established that "decision rules which do not allow any compensation at all tend to show the worst results in terms of classification accuracy" (Van Rijn et al, 2012). This is presumably the rationale for 'grace marks', 5 marks out of 550 deployed over two subjects at the pass/fail borderline to the candidate's best advantage. Otherwise standards are absolute and absolutely the standards of the BISE. The candidates play no obvious part in establishing standards. The cut-offs are determined before a single candidate is examined and will remain fixed until IBCC chooses to change them.

How different are O level standards. They are essentially subject based and as subjects are intrinsically different fall back upon man, or at least boys and girls, as the measure of all things. The O level standard was initially that about 60% of about 15% of the population who stayed on beyond the school leaving age, should be awarded a pass. So O level standards were based on percentages of candidates, not percentages of marks. It is the candidates and of course their schools who set the standard in "a socio-cultural paradigm [which] recognises assessment as a context-dependent, socially-situated, interpretative activity" (Shay, 2004). No wonder that where parents can afford it schools flock to O level. When CIE approached IBCC for recognition of its new A* grade as equivalent to 95% of available marks, CIE recognised the inherent improbability of such a mark in any subject which leaves room for examiners' judgement with the reassurance that this would be a rare award to a few outstanding individuals. But in the socio-cultural paradigm, the grades are in fact examiners' judgements of what the marks may mean. It sometimes seems that in Pakistan's private schools it is grades other than A* and A which are rare awards these days but we do not know. Contrary to the guidelines laid down in *Standards for Educational and Psychological Testing* (1999) which encapsulates the combined wisdom of the American Educational Research Association, the American Psychological Association and the National Council for Measurement in Education, CIE does not publish its Pakistani grade or score distributions. It does not reveal the social impact of its grading decisions. They are a commercial secret.

There are thus two approaches to standard setting in Pakistan, standards defined by the questions and their mark schemes which are determined in advance and standards defined by the examiners who ensure that the best candidates' performances are deemed worthy of 95% of marks regardless of subject. The Cambridge approach betrays its origins in norm –referencing, whereas the BISE are operating a kind of absolutist criterion referencing: a competence is demonstrated or it is not, and that is that. The BISE standards are embedded in the construction of the paper.

The educational impact of the two examination systems can be evaluated by reference to AKU's University Admission Test but only for a self-selected group of high flying candidates. Candidates know their SSC results when they make the decision to apply. Most of them will also have an impression that to get into AKU medicine you have to be a pretty high achiever, but the strength of that impression is not known and is certainly not uniform given that applicants are distributed nationwide. The conventional approach to equating results through a common calibrating instrument is to scale each Board's distribution of scores using the means and sds of the calibrating test, in this case the UAT, but the accuracy of that procedure depends upon the representativeness of the Board samples. If the students attempting UAT are a biased sample of the students from the parent board and the bias is not a constant across boards, the measures of central tendency upon which such comparisons are based will be differently centred, rendering the comparison untenable.

Competitive edge

In seeking an alternative to scaling, we return to the notion of "being in with a chance". The various boards have been compared by considering whether they confer any competitive advantage on their students, that is whether students at the same grade level experience different fates. CIE with a top grade equivalent, they claim, to 95% of marks occupies a space for which the BISE do not even have a name. Their A1 grade starts at 80% of marks though the Federal Government of Pakistan makes a cash award to every candidate who scores more than 90% SSC marks. Given the prevalence of O level students in the applicant samples, the analysis has had to play in O level territory. Candidates recorded as scoring between 90% and 100% and between 80% and 89.9% are located in 'bounded' categories, that is the top and bottom of the grades is known and what lies between the cut-offs shares a common designation in all BISE. While we cannot evaluate the general educational standards of the BISE by reference to UAT we can establish whether any board gives talented candidates receiving the same overall SSC grade a competitive edge in pursuing an academic education in medicine.

Table-3 sets out the odds on being considered for an MBBS place for candidates with SSC scores in the same grade interval. Not included are 52 applicants from the US of whom 4 were shortlisted, nor 43 applicants, none successful, taking school examinations from other countries.

There are four noteworthy features of Table-3 in relation to public examination standards in Pakistan. The first is the consistency in the pattern of outcomes between years. In every board the examination questions are different in consecutive years but the educational value of its top grades in the premedical group remains remarkably stable even though the number of top grade candidates applying for AKU fluctuates quite widely. There is clear evidence that a standard is being maintained in consecutive years, but unfortunately it is not the same standard in every board.

Karachi and the Sindh BISE are at one extreme. Not a single applicant has reached the shortlist in 2011 and only two in 2012. The Sindh boards examine half of the elective subjects in Year IX and the remaining subjects in year X. All other Pakistani Boards examine every subject twice, covering half of the subject syllabus in year IX

and the rest of the subject syllabus in year X. Sindh has never offered a rationale for not falling into line with the other provinces but the implicit message that subjects are bodies of knowledge to be mugged up just for the short term is associated with extremely deleterious wash back effects. In each of the three sciences the Karachi mean UAT score is more than half a standard deviation below the overall mean of all applicants and in science reasoning it falls more than a full standard deviation short of the overall mean, a huge effect. It is not that there is no talent in the whole of Sindh. These young people have been misled by the apparent value of their BISE marks.

At the other extreme we have O levels which certainly dominate the pursuit of places. The effect of grading as an interpretative activity (Christie and Forrest,1981; Gipps,1999) is clear to see. O level borderlines are drawn by very senior examiners who have educational values, by no means always venal. They clearly have a view of A* as having much less to do with 95% than with potential to benefit from a high quality university education. It is notoriously difficult to capture that potential in a verbal description but they seem to share much the same assumptions that underlie AKU's UAT. The O level examiners are spotting talent that has flowered into remarkable educational achievement in very privileged schools and they are good at it. Look at the fate in 2012 of those they have passed over from the same privileged schools: it is no different from that of high grade SSC candidates who have had one fewer year of rather more mundane schooling. This is not the educational diet for everyone.

> "A deep strategic approach to studying is generally related to high levels of academic achievement, but only where the assessment procedures emphasize and reward personal understanding. Otherwise surface strategic approaches may well prove more adaptive." (Entwistle, 2000, p4)

That leaves the fourth feature of Table-3, the noise in between which is gradually taking on a more structured appearance. The two BISE with a national remit, the Federal Board and AKU-EB, are at the moment clearly distinguished from the BISE with a provincial identity. The distinction is at least in part curricular. Examining across provinces the national boards have to examine not text books but syllabuses, the 2006 revision of the National Curriculum in both cases. Its emphasis on understanding and the application of knowledge is clearly paying dividends in the expansion of genuine educational opportunity beyond the deep pockets required for O level schooling. But the distinction is not permanent. The Lahore BISE is just two years behind them in adopting the 2006 syllabuses and the benefit is being further reinforced by the reorganisation of the Punjab BISE. They have adopted a new strategy of giving each Board sole responsibility for a few subjects for the whole province. The emergence of a subject focus within the Punjab Boards is an important counterpoise to the "one strategy fits all" that underlies rote memorisation. If standards are associated with subjects rather than an overall aggregate, distinctive features of subject performance come into play and examiners begin to look for subject appreciation rather than excellent recall. There is some evidence that such a shift is occurring in Lahore. The top tier are getting a sound preparation for university education. There is a qualitative difference between these students and those who fall in the 80 – 89% grade. The hope for education

in Pakistan is that as the exam boards get better at identifying the skills which are relevant for higher education the students they identify will be able to sustain their performance levels for the benefit of the nation as a whole.

References

Bloxham, S and P, Boyd (2012), Accountability in grading student work: securing academic standards on a Twenty-first century quality assurance context, British Education Research Journal, 38, 4, p615-634.

Christie, T. And Forrest, G.M. (1981), Defining public examination standards, London Macmillan Education

Gipps, C (1999) Socio-cultural aspects of assessment. Review of Research in Education 24, 355-392.

Shay, SB (2004), The assessment of complex performance: a socially situated interpretative act, Harvard Education Review 74 (3), p307-329.

Van Rijn, PW, Beguin AA and HHFM Verstralen (2012), Issues and implications of high stakes decision making in final exams in secondary education in the Netherlands, Assessment in Education Principles, Policy and Practice 19(1), p117-136.

Table-1: Internal structure of the AKU University Admission Test

(Correlations above and to the right of the diagonal relate to 2011, below and to the left to 2012.)

Overall (n=4273)	**Eng. Essay**	**Eng. MCQ**	**Biology**	**Chemistry**	**Physics**	**Science** Reasoning	**Maths** Reasoning	**2011 Total**
Eng. Essay (18 marks)		0.46	0.41	0.30	0.21	0.46	0.38	0.61
Eng. MCQ (30 Marks)	0.38		0.55	0.32	0.24	0.56	0.44	0.73
Biology (20 marks)	0.31	0.58		0.60	0.51	0.62	0.41	0.82
Chemistry (20 marks)	0.24	0.31	0.55		0.63	0.46	0.24	0.70
Physics (20 marks)	0.26	0.39	0.58	0.62		0.37	0.18	0.60
Science Reasoning (30 marks)	0.32	0.53	0.69	0.62	0.61		0.60	0.84
Maths Reasoning (20 marks)	0.33	0.59	0.56	0.31	0.42	0.54		0.66
2012 Total (158 marks)	**0.48**	**0.75**	**0.84**	**0.70**	**0.75**	**0.85**	**0.75**	

Table-2: First principal component of UAT in 2011 and in 2012.

	2011 (n=4879)			2012 (n=3196)		
Subject						
	Marks	**Mean**	**S.D.**	**Factor 1**	**Mean**	**S.D.**	**Factor 1**
English Essay	18	9.23	3.30	.64	9.99	2.57	.50
English Reading	30	14.08	4.68	.73	17.01	5.78	.73
Biology	20	9.66	4.65	.84	9.91	5.22	.84
Chemistry	20	9.05	4.63	.71	7.82	4.79	.72
Physics	20	5.91	3.52	.61	7.86	4.54	.76
Science Reasoning	30	16.43	5.73	.85	17.10	5.47	.85
Maths Reasoning	20	8.67	4.49	.70	12.86	5.82	.70
Total Score	**158**	**63.79**	**20.56**		**82.55**	**25.74**	
% Variance explained				**54%**			**53%**

Table-3: Probability of being included in the top 10 percent of applicants to AKU's MBBS of the top two grades of school achievement scores.

	School marks 90 – 100%					**School marks 80 – 89.9%**			
BISE									
	2011		2012			2011		2012	
	n	*p*	n	*p*		n	*p*	n	*p*
AKU-EB	*5*	*0.20*	*3*	*0.33*		58	*0.10*	54	*0.07*
FBISE	*67*	*0.16*	*73*	*0.19*		213	*0.05*	142	*0.06*
BISE Lahore	220	*0.11*	89	*0.06*		144	*0.01*	71	*0.00*
Rest of Punjab	*332*	*0.06*	*133*	*0.11*		390	*0.02*	102	*0.02*
All KPK,AJK G-B	82	*0.05*	79	*0.08*		439	*0.01*	498	*0.01*
BISE Karachi	*23*	*0.00*	*5*	*0.00*		219	*0.00*	140	*0.01*
Rest of Sindh	10	*0.00*	14	*0.00*		128	*0.00*	115	*0.00*
All BISE	*739*	*0.08*	*396*	*0.10*		1591	*0.02*	1122	*0.02*
O Levels	902	*0.37*	873	*0.28*		218	*0.10*	177	*0.03*

31 October 2012

Establishing standards for grading (checklists and rubrics)

Developed by Dr Thomas Christie

Brainstorming

Ask Student Teachers about their ideas on criteria for grading. As they respond, write their thoughts on the board. This will allow you to gauge their level of knowledge and organize their ideas. After the Student Teachers have finished making suggestions, analyse their ideas in a presentation.

Instructor presentation

Introduce standards for grading. Explain the criteria for grading and the three types of criteria: the nature and number of assessments, the weight given to each assessment, and performance standards. Provide an illustration of each.

Q and A session

Check Student Teachers' understanding of the material by asking the following questions:

- How is it possible to determine if grades provide information on the achievement of course objectives?
- What are four factors used in establishing performance standards?

Responses to the second question should address the following points:

- The grading process should be unbiased and compare each student to the same criteria.
- Grades should be based on sufficient data to permit valid evaluations of students' achievement to be made.
- The component items should be related to instructional objectives.
- The grade names should be generally understood (e.g. pass/fail; first, second, third; A, B, C).

How often to grade?

There are no hard and fast rules. However, to determine an appropriate time, it is often worthwhile to look at the difference in rate of learning between the highest achievers and lowest achievers. When the time comes to move to a new topic, most teachers focus on two or three pupils at the top of the bottom quartile, the class's 20th percentile. Once this group shows signs of mastering the topic, the high achievers have likely become bored and restless, and it is time for a new topic. This has been formalized as mastery learning (Block and Airasian, 1971, pp. 37–38). The teacher introduces new material, covers it for about two weeks, and then tests the class. When the results are returned, each non-master is paired with a master for peer teaching. Those in between fend for themselves or work in groups.

After three weeks, the whole class can move on. The mastery learning approach usually establishes a pattern that leads to at least four classroom tests in a term. If each test has a reliability of 0.8, then according to the Spearman–Brown prophecy formula, the reliability of the aggregate score would be 0.93, which is as high as achievement test reliabilities get.

That calculation brings us as close to a rule for the number of grades we will get. The rule says that while it is inevitable that a grade award may be wrong by one grade, it should only very rarely be wrong by two grades. To make sure we are within these limits, calculate the standard error of measurement, SE mean = SD$\sqrt{1 - r_{tt}}$. If the correlation coefficient is $r_{tt} = 0.93$, then the SE is a little over a quarter of the SD. The SE mean is assumed to be normally distributed, so in the highest risk case, where the observed score is the grade cut-off score, the next cut-off would have to be more than two standard errors away, or 0.6 standard deviations. Even a very good test will have a range of only six standard deviations. Thus, the maximum number of grades in these ideal conditions is 10. Half that number would be safer. That still leaves a further problem near the mean. If you choose five grades, more than half of the examinees will get the same grade, grade three on a five-point scale. If you choose to have six grades, you have to make difficult decisions where the greatest number of people will likely be affected by error (see the normal curve). It becomes a question of whether you are trying to build achievement for most or success for some.

Interpretation of test scores

Developed by Samreen Jalal

Brainstorming

Ask Student Teachers about their ideas about the methods of interpreting test scores. As they respond, write their thoughts on the board. This will allow you to gauge their level of knowledge and organize their ideas. After the Student Teachers have finished making suggestions, analyse their ideas in a presentation.

Instructor presentation

Introduce, explain, and provide specific examples of methods of interpreting test scores. Discuss criterion-referenced methods related to the grading standards used by the National Curriculum 2006 revision and norm-referenced methods, especially marking on the curve, which deliberately distributes marks to conform to the areas under the normal curve.

Think, pair, share

Divide Student Teachers in pairs. Ask them to think about potential pitfalls in interpreting test scores.

Instructor presentation

Introduce and explain the issues in interpreting test scores.

Consider using the following references in creating the presentation for this topic:

- Shermis, M, D., & Divesta, F, J. (2011). *Classroom assessment in action* (pp. 28–34). Lanham, MD: Rowman & Littlefield Publishers.
- Miller, M. D., Linn, R. L,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.) (pp. 461–465, 486–489). Chicago, IL: R. R. Donnelley & Sons Co.

Modes of reporting

Developed by Samreen Jalal

Brainstorming

Ask Student Teachers about their ideas about modes of reporting. As they respond, write their thoughts on the board. This will allow you to gauge their level of knowledge and organize their ideas. After the Student Teachers have finished making suggestions, analyse their ideas in a presentation.

Q and A session

Ask Student Teachers the following questions:

- Which method is best for clarifying a student's progress in learning?
- What is used to maintain a record of progress that follows a student from grade to grade?
- What is used to provide a basis of discussion with parents?

Reporting assessment results and problems in grading and reporting

Developed by Samreen Jalal

Examples of student report cards are required for this session.

Brainstorming

Ask Student Teachers to share their ideas about reporting assessment results. As they respond, write their thoughts on the board. This will allow you to gauge their level of knowledge and organize their ideas. After the Student Teachers have finished making suggestions, analyse their ideas in a presentation.

Q and A session

Ask Student Teachers the following questions:

- What is the best way to communicate to students and parents a student's performance in college and potential for further success?
- What are some ways to organize a lesson, unit, or semester in order to mark transitions in a course, including conclusions?

Instructor presentation

Introduce issues related to reporting assessment results and provide examples related to each. Discuss this in relation to reporting results to students, other teachers, parents, and the school, and reference problems in grading and reporting. Consider using teaching aids, such as charts, an overhead projector, or a PowerPoint presentation, to supplement and illustrate your presentation.

One-minute paper

Ask Student Teachers to write a one-minute paper on the ideas and concepts covered in the Instructor presentation.

Instructor presentation

Introduce uses and modes of reporting. For students, report card grades should not only enable them to interpret their progress since their last report, but they should also indicate in what ways progress has been made and what remediation is necessary to get back on track. For parents, a report card provides an accurate picture of their child's strengths and weaknesses. It can inform what they need to do to help their child and in what areas remedial help is needed. In parent–teacher conferences, a well-considered report of student progress provides reasonable justification for an assigned grade.

Presenting data

Highlight the common features and fields on student report cards and share examples. Student Teachers will share the features that were present on their own report cards at school.

The discussion should address what messages report cards convey about education in the institution in which they are used. Student Teachers should also consider whether report cards are appropriate, how they might be improved, and why.

Conclusion

Conclude the lesson with a summary of the session and answer any questions that Student Teachers may have.

The following resources may be helpful in summarizing the lesson or creating the actual lesson:

Department of Education and Early Childhood Development (n.d.). Sample report cards. Retrieved from:

- http://www.education.vic.gov.au/school/teachers/support/pages/reportcards.aspx?Redirect=1

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.) (pp. 265–267). New Delhi: Prentice Hall of India.

Education Review Office (2008). Assessment in primary schools: A guide for parents. Retrieved from:

- http://www.ero.govt.nz/National-Reports/Assessment-in-Primary-Schools-A-Guide-for-Parents-December-2008

Guskey, T. R. (2010). *Grading and reporting student learning* (pp. 2–8). Thousand Oaks, CA: Sage Publications.

Marzano, R. J. (2006). *Classroom assessment and grading that work* (pp. 125–132). Alexandria, VA: Association of Supervision and Curriculum Development.

Shermis, M, D., & Divesta, F. J. (2011). *Classroom assessment in action* (pp. 361–365). Lanham, MD: Rowman & Littlefield Publishers.

Walvoord, B, E., & Anderson, V. J. (1998). *Effective grading: A tool for learning and assessment* (pp. 2–5, p. 125). San Francisco: Jossey-Bass.

EVALUATION

Key words

- Classroom observation
- Observation format
- Observation results
- Portfolio
- Types of portfolios
- Portfolio assessment process
- Oral questioning
- Anecdotal records and notes
- Student judgements and reports
- Student progress and teaching-learning process
- School effectiveness and improvement

Classroom observation

Developed by Dr Bashir Hussain and Dr Shafqat Hussain

Observation exercise

Ask the Student Teachers to work in pairs on an observation exercise. Each pair will leave class for five minutes and observe their surroundings. The pairs will come back and report two things: the focus (e.g. the whole scene, the building's physical structure, people) and what specifically was observed (e.g. the building has five rooms with wide glass windows).

After all the Student Teachers have reported their observations, sum up the activity by explaining that the observation was an assessment. This form of assessment is used for ongoing assessment of teaching and learning taking place in the classroom. Teachers observe students in class to gauge a variety of things, such as who in class is excited, bored, frustrated, motivated, or making progress. Successful teachers know how to use these observations to modify their teaching and mould student learning.

Purpose and principles of classroom observation

Distribute and discuss the handout 'Classroom observation protocol', available from http://www.snct.org.uk/library/1231/Classroom%20Observation%20Protocol.pdf, which discusses the purpose and principles of classroom observation.

Classroom observation to promote assessment for learning

Share the 10 principles from Gardner's *Assessment for learning.* Ask Student Teachers to identify any three principles that are directly fulfilled by effective classroom observation.

Assessment for learning

- is a part of planning;
- focuses on how students learn;
- is essential to classroom practice;
- is a key professional tool;
- is sensitive and constructive;
- fosters motivation;
- promotes understanding of goals and criteria;
- helps learners know how to improve;
- develops the capacity for self-assessment;
- recognizes all educational achievement.

From Gardner, J. (2006). *Assessment and learning: An introduction* (p. 3). Gateshead, UK: Sage Publications.

Ask each Student Teacher to share their chosen principles. Then have the class engage in a debate to come to a consensus on three to five common principles that are directly fulfilled by classroom observation.

Planning and preparing for observations and the typical observation format

Developed by Dr Bashir Hussain and Dr Shafqat Hussain

Requirements for classroom observation

Explain that when classroom observation is used as an assessment method, it requires careful planning by the observer. A concise list of planning requirements is provided in 'Classroom observation' (available from http://www.atl.org.uk/Images/ADV19%20Classroom%20observation.pdf). Print this document and use it as a handout.

What to observe

Remind Student Teachers of the observation exercise they completed and how their focus determined what they observed. As they were not able to observe everything in their surroundings during the exercise, they will not be able to do so in a classroom either. It is important to determine an area of focus in teaching and learning. This focus informs who will be observed – the teacher, the students, or a few students. In addition, the focus determines which aspects of teaching and learning will be observed, such as teacher attention, teacher–student interaction, student–student interaction, and student communication.

Divide Student Teachers into groups of six and ask each group to select three features to be the focus of an observation in a primary classroom. Before choosing the features, the group needs to decide the purpose of observation and who to observe in the class.

Share the following observation tools, which will provide the Student Teachers with some guidance:

- Classroom observation checklist (n.d.). Available from:
  - http://www.gadsdenstate.edu/faculty-and-staff/ie/documents/ClassroomObservationChecklist.pdf
- Classroom observation checklist (n.d.). Available from:
  - http://www.unl.edu/gradstudies/current/teaching/Classroom_Observation_Form.pdf
- Classroom observation checklist (n.d.). Available from:
  - http://www.austincc.edu/hr/eval/procedures/ClassObservCheck.pdf

Homework

Each group will develop an observation form, and each group member will go to two primary school classrooms and record observations.

How to derive results from the observation

Developed by Isbah Mustafa

Compiling and deriving results from the observation

The groups will compile their observations from the primary school classrooms. For binary or scale observation, frequency will be used. For observation by frequency of occurrence, the groups will use mean and standard deviation as the basic tool for making meaning of their observations.

Each group will derive results from the observation data. They should look for outliers while analysing the data to guard against any misinterpretation. Each group will present their findings to the class. Terms such as *satisfactory, unsatisfactory,* and *proficient* can be used to present the findings.

Homework

Each group will write a report on their classroom observation. The report will contain the information about the focus, observation tool, the number of observations, and major findings. The report will be included as part of the final assessment.

Oral questioning

Share data on teacher questioning behaviour based on observations of a sample of government middle school classrooms in Pakistan. Ask Student Teachers to interpret the following data from 1058 classrooms in government middle schools, and then discuss oral questioning practices in primary schools in Pakistan.

Questioning style of the teacher	(in %)
The teacher asks more than two questions	77
The teacher asks open-ended question	44
The teacher gives adequate response time	34
The teacher gives explanatory comments	44
The teacher checks for understanding of the topic	47
The teacher takes response from a selected group of students	13

Source: Mustafa, I. (2010) EDLINKS close out report by Aga Khan University Examination Board

Effective questioning

Effective questioning involves the careful framing of questions, wait time, the involvement of the majority of the students, peer discussion by students to reach to an answer, open-ended and problem-solving questions, positive reinforcement, reflection by the teacher, and teacher feedback (Black, Harrison, Lee, Marshall, and Wiliam, 2003; Clarke, 2005).

Discuss each of the elements and encourage the Student Teachers to think of the impact each element has on teaching and learning.

Conclusion

Student Teachers will write a paragraph on how questioning in primary classrooms can be made effective in Pakistan.

Student portfolios

Developed by Zahid Majeed

Introduction

Ask Student Teachers the following questions on student portfolios, but do not mention the lecture topic:

- How have your Instructors evaluated you progress during your university studies?
- Can you highlight the main elements of evaluation criteria used by your university Instructors?
- What different tools were used to evaluate your progress? Why were these tools used?
- What procedure did the university use to inform you about your progress in a particular subject or semester?

After the Student Teachers have responded to these questions and discussed the related issues, introduce the topic of this session: student portfolios.

Student portfolios

Briefly describe student portfolios and provide a definition. Make sure that Student Teachers understand the difference between progress and performance.

Portfolio evaluation criteria

Give a brief lecture on measuring student progress through the use of a portfolio. Share some examples of evaluation criteria.

Group work

Supply four or five student portfolios and rubrics that can be used to evaluate them. Portfolios can be downloaded from:

- http://sbs.mnsu.edu/socialstudies/studentportfolios/evaluate.html.

Divide Student Teachers into groups of four or five and have them evaluate the portfolios based on the rubrics provided. The Student Teachers will share their assessment of student progress through portfolios. Each group will present their evaluations to the class.

Conclusion

Summarize the features and purpose of student portfolios and the related evaluation process.

Homework

Ask Student Teachers to develop an outline of the contents of a student portfolio. A portfolio intends to measure student performance over a semester. Student Teachers should also try to develop criteria to evaluate the portfolio.

References

Brennan, R. L. (2006). *Educational measurement*. Westport, CT: Praeger Publishers.

Miller, M. D., Linn, R. L,. & Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed.) (pp. 461–465, 486–489). Upper Saddle River, NJ: R. R. Donnelley & Sons Co.

Portfolio evaluation rubric (2012). Retrieved from:

- http://www.uvm.edu/~cnhs/resources/CSD_portfolio_evaluation_form.pdf

Portfolio rubric (n.d.). Retrieved from :

- http://www.aug.edu/~tdevwww/faculty_info/Initial_Ceritification_Portfolio_Rubric.pdf

Prince George's County Public Schools (n.d.). Portfolio assessment. Retrieved from:

- http://www.pgcps.org/~elc/portfolio.html

Anecdotal records, student judgements, and reports

Developed by Zahid Majeed

Brainstorming

Recall the discussion on student portfolios with Student Teachers, and ask the following questions:

- What tools are available to assess student achievement and learning?
- How are these tools used to assess the students' academic achievements?

Introduce the session topic: different assessment tools used by primary school teachers to evaluate student performance and achievement.

Print 'Student assessment' (available from http://learndat.tech.msu.edu/teach/student-assessment) and use it as a reading. Ask Student Teachers what they learned from this reading, and engage them in a whole-class discussion.

Anecdotal report

Explain the tools used to assess student learning and achievements. Share anecdotal record-keeping devices. Samples are available from:

- http://learndat.tech.msu.edu/teach/student-assessment.

Divide Student Teachers into groups of four or five, and ask them to decide which of these tools can be used to assess student learning in government primary schools. Each group will share their points with the group sitting next to them.

Student judgements

Ask Student Teachers to think how student judgement, in the form of self- and peer assessments, can be incorporated into student assessments. Responses should reference techniques such as peer marking of student work in summative examinations.

Conclusion

Summarize the main points of the session.

Assessment

Ask Student Teachers to visit a primary school near their home and interview a teacher on the tools used to assess student progress and how those tools are used.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.) (chapter 17). New York: Macmillan.

Copeland, C. (2013). Assessment tools to monitor students' progress. Retrieved from:

- http://www.ehow.com/list_7710789_assessment-tools-monitor-students-progress.html

Elrod, J. (n.d.). Assessment tools used in schools. Retrieved from:

- http://www.ehow.com/info_7956483_assessment-tools-used-schools.html

Leo, J. (n.d.). Teachers' assessment tools. Retrieved from:

- http://www.ehow.com/info_7903589_teachers-assessment-tools.html

Mehrens, W. A., & Lehmann, I. J. (1987). *Using standardized tests in education* (4th ed.) (chapter 9). New York: Longman.

Miller, M. D., Linn, R. L. & Gronlund, N. E. (2009). *Measurement and assessment in teaching.* New Jersey: R. R. Donnelley & Sons Company.

Student assessment (n.d.). Retrieved from:

- http://learndat.tech.msu.edu/teach/student-assessment

Analysing and reporting student progress:

Developed by Zahid Majeed

Introduction

Review the previous discussion on student portfolios. Note that summative examinations can be used for some of the same purposes as a portfolio. Longitudinal data about student performance within a cohort indicates progress made by the group and individuals over time.

Use 'Test scores: A guide to understanding and using test results' by D. P. Flanagan and L. F. Caltabiano, available from http://www.nasponline.org/communications/spawareness/testscores.pdf, to prepare a presentation on analysing and reporting student progress.

Prerequisites for measuring student progress
Ask Student Teachers to think about what is needed in a summative examination to inform a teacher of student progress. Answers should include:

- standards-referenced tests
- standardized marking
- uniformity in reporting student scores.

Issues in measuring student progress using raw scores

Share copies of used mark sheets from SSC and HSSC exams. In pairs, Student Teachers will compare marks from the two results to evaluate student progress.

Each pair will work with another pair to compile a list of potential problems and limitations in using student scores to measure progress. This list will be shared with the class.

Conclusion

Summarize the challenges in using scores to measure progress and, in particular, note the lack of standardized marking and the limited use of raw scores in reporting in Pakistan.

Homework

Student Teachers will write a paragraph on measures required in government primary schools in Pakistan to produce student achievement data that measures student progress.

School effectiveness and classroom improvement

Developed by Zahid Majeed

Introduction

Define the terms *school effectiveness* and *classroom improvement*. These ideas are based on the assumption that schools influence student development. Specifically, financial and human resources, context, and process are the inputs that influence student development. The classroom improvement model does not compare input and output; rather it compares two outputs from the same institution to gauge improvement. The most valuable output is information about student achievement.

Brainstorming

Working in groups, Student Teachers will brainstorm how student achievement data can reflect school effectiveness and classroom improvement.

The groups will then come together to discuss their thoughts with the class. As they respond, write their thoughts on the board. This will allow you to gauge their level of knowledge and organize their ideas. After the Student Teachers have finished making suggestions, analyse their ideas in a presentation.

School improvement life cycle

To introduce the subject, use the school improvement life cycle, available from http://www.advanc-ed.org/school-improvement-life-cycle. The cycle involves using achievement data to help set goals for schools, design efforts to achieve those goals, and evaluate the efforts.

A model of school improvement

To introduce the model, use the study by Mourshed, Chikioke, and Barber, 'How the world's most improved school systems keep getting better' (available from https://mckinseyonsociety.com/downloads/reports/Education/How-the-Worlds-Most-Improved-School-Systems-Keep-Getting-Better_Download-version_Final.pdf). Distribute a copy of page 30 to the class. This page summarizes the efforts identified as effective in moving a school system from poor to fair.

After Student Teachers have read the summary, ask them to identify three efforts from the list that would be most effective in attempts to improve government primary schools in Pakistan. After each Student Teacher shares his or her preferences, the class will debate to reach a consensus on which three interventions can be most relevant and effective in the Pakistani context.

Conclusion

Recap the definitions of school effectiveness and classroom improvement as well as the use of student achievement data to gauge improvement and effectiveness.

5

Methods and strategies to use in planning teaching and learning

The following is a list of some of the strategies used in this course to encourage active learning.

Active lecturing. An active lecture is not too different from any good lecture, but it attempts to directly involve listeners.

There is no one best way to give an active lecture, but it involves any of the following techniques

Give information in small chunks (about 10 minutes), and then have class members do something with that information for a few minutes. Here are some examples of activities, which you can repeat or vary:

- Write a one-minute reaction to what you have just heard.
    - Talk to the person next to you about what you heard and see how your perspectives differ. Do you agree? Do you have questions?
- List as many key points as you can remember.
- Compare notes taken during the chunk. Help each other fill in gaps or determine if crucial information is missing. (Some people do not allow note taking during the lecture, but this is up to the Instructor.)

Give out cards or slips of paper in three different colours. When class members are listening to your comments, have them hold up a colour for 'I understand', 'I don't understand', or 'I disagree'. Then either stop and allow questions, or adjust what you are saying so there are more 'understand' colours showing. This is particularly effective with large groups of 50 or more people.

Ambassadors. This is a useful way to get groups or individuals to exchange information. Two or more members move from one group to another to share/compare discussion etc. You may wish to have half of each group move to another group. This is especially useful if you do not have ample time for a whole-class discussion.

Brainstorming. This is a technique for generating creative ideas on a topic. It may be an individual activity or organized as a group activity. Give people a limited amount of time (e.g. one minute) to say or write as many ideas as they can on a topic. No matter how unrelated an idea seems, write it down. (Alternatively, the Instructor might ask the whole class to brainstorm and write all the ideas on the board.) After the brief period of brainstorming, ideas may then be analysed, organized, and discussed. This is often used as a problem-solving technique. Ideas are then analysed in light of how useful they might be in solving the problem.

Gallery walk. This is a strategy that borrows its name from a visit to an art gallery. Students walk through an exhibit of posters, artefacts, or display of items they have completed. They can be directed to take notes. The idea is to thoughtfully look at what is displayed.

Graffiti wall. A graffiti wall may be displayed in the classroom for use all term. Students may write their thoughts, feelings, or expressions before or following each session and sign their name. Anonymous comments are not suitable. Ideas generated

in class may be posted on the 'wall'. Use paper from a large roll of craft or newsprint paper or join several cardboard boxes together to make a wall that can be stored between sessions. Students can take turns getting and putting away the wall each session.

Group work: some tips for forming instructional groups. There is no one best way to form groups. The best way for you is the way that suits your purpose. Use a more complicated strategy if students need a break or need to be energized. Use a simple technique if time is short. Ways to form groups include the following:

- Ask people to count off from one to five (depending on the number of people you want in a group). Groups will form based on their number (e.g. all of the ones will gather together).
- Before class, determine how many people you want in a group or how many groups you need. Give each class member a different coloured sticker, star, or dot as they enter the class. Then when it is time to form groups, ask them to find people with the same sticker etc. and sit together.
- Put different coloured bits of paper in a cup or jar on each table. Have people take one and find people in the room with the same colour to form a group.
- Have students get together with everybody born in the same month as they were.

Make adjustments to the groups as needed.

Mini-lecture. A mini-lecture contains all the components of a good lecture. It is sharply focused. It begins with an introduction that provides an overview of what you will talk about.

It offers examples and illustrations of each point. It concludes with a summary of the main point(s).

One-minute paper. Ask class members to write for one minute on a particular topic (e.g. their reflections on a topic, an assigned subject). They are to focus on writing their ideas, without worrying about grammar and spelling. A one-minute paper differs from brainstorming because there is more focus.

Pair-share. Use this technique when you want two class members to work together to share ideas or accomplish a task. Simply ask them to work with a neighbour or have them find a partner based on some other criteria. It is very useful when you want people to quickly exchange ideas without disrupting the flow of the class. (Sharing in triads and foursomes are also small group techniques.)

Poster session. This is useful when you want students to organize their thoughts on a topic and present it to others in a quick but focused way. Have individuals or small groups work to create a poster to explain or describe something. For example, if they have been doing an inquiry on a particular topic, they would want to include their focus, methods, and outcomes, along with colourful illustrations or photographs. The poster can be self-explanatory or students can use it to explain their work. As an in-class tool, a poster session is often combined with a gallery walk so that the class may review a number of posters in a short time.

Readers' theatre. readers' theatre is a group dramatic reading from a text. Readers take turns reading all or parts of a passage. The focus is on oral expression of the part being read rather than on acting and costumes. readers' theatre is a way to bring a text to life.

It is a good idea to go over passages to be read aloud with students so they are familiar with any difficult words.

Sometimes readers' theatre is used to get student interested in a text. They hear passages read first and then read the longer text.

KWL. This is a strategy that provides a structure for recalling what students know (K) about a topic, noting what students want to know (W), and finally listing what has already been learned and is yet to be learned (L).

The KWL strategy allows students to take inventory of what they already know and what they want to know. Students can categorize information about the topic that they expect to use as they progress through a lesson or unit.

Text-against-text. This is a way of helping students learn to analyse and compare written documents. The idea is to look at two documents and search for overlap, confirmation, or disagreement. It is a way of looking at different perspectives. Sometimes it is useful to give students readings prior to class and ask them to compare the readings based on a set of study questions, such as:

1) Look at each author separately. What do you think the author's main point is?
2) How does the author support his/her argument?
3) Look at the authors together. In what ways do the authors agree?
4) What are their points of disagreement?
5) What is your opinion on the issue?

Text-against-text may be used to compare a new reading or new information with material that has already been covered.

In classrooms where the whole class uses a single textbook, Instructors often find they are teaching against what is in the textbook. Sometimes it is hard for students to accept that a textbook can and should be questioned. Putting together a text against text activity using the textbook and outside materials (e.g. an article) can help them understand that there are legitimate differences of opinion on a subject. Articles need not contradict each other. They may be about the same topic, but offer students different ways of seeing a subject.

Another way to use the activity is divide the class into groups, give each a set of materials, and have them debate the texts. Some university faculty like to put together text sets that include both scholarly and non-scholarly works and have students to think about differences. For example, you might provide all students—regardless of their reading level or learning style—with easy-to-read materials as a way to introduce themselves to a topic. Even competent adult learners seek out 'easy' books or materials to learn about a new or complex topic. Providing a picture, newspaper article,

or even a children's book in a text set might give everyone the means of connecting to or understanding some aspect of the larger subject.

Roundtable technique. For this technique, divide the class into small groups (i.e. four to six people), with one person appointed as the recorder. A question that has many possible answers is posed, and class members are given time to think about the answers. After the thinking period, members of the team share their responses with one another. The recorder writes the group's answers. The person next to the recorder starts and each person in the group (in order) gives an answer until time is called.

Quizzes. Prepare and give a short quiz (15 minutes) over the different aspects of test development and evaluation covered in the unit. As students take the quiz, ask them to circle items they are unsure of. They can review and discuss their work in the following ways:

- *Triads*. Have students meet in groups of three to review the quizzes so that they can help each other with their weak areas. (10 minutes)
- *Review.* Go over the quiz with students, and have them look at their own work and make corrections. (30 minutes)
    - Notice points class members had difficulty remembering and take time to review them. You may ask students to assist with this and discuss how they were able to remember.
    - Use this time to correct any misconceptions.
    - Have students save their quiz for future study.

Resources

Accountability

Hout, M., & Elliott, S. W. (2011). Test as performance measure. In *Incentive and test-based accountability in education* (pp. 37–52). Washington, DC: The National Academies Press. Available from:

- http://www.nap.edu/openbook.php?record_id=12521&page=37

Adams, S. J., Heywood, J.S., & Rothstein, R. (2009). *Teachers, performance pay, and accountability: What education should learn from other sectors.* Washington, DC: Economic Policy Institute. Retrieved from:

- http://www.epi.org/page/-/pdf/teachers_performance_pay_and_accountability-full_text.pdf

Alpha and Beta testing by study mode

jmgallegos78 (2011). Alpha and Beta testing. Retrieved from:

- http://www.studymode.com/essays/Alpha-And-Beta-Testing-669067.html

The Alpha and Beta tests of World War I (n.d.). Retrieved from:

- http://www.nald.ca/library/research/adlitus/page19.htm

Assessment

Educational assessment (n.d.). Retrieved 27 June 2013 from:

- http://en.wikipedia.org/wiki/Educational_assessment

What can I learn about assessment from this website? (n.d.) Retrieved from:

- http://www.cmu.edu/teaching/assessment/

Authenticity, objectivity, adequacy, and differentiability

Lewkowicz, J. A. (2000). Authenticity in language testing. *Language Testing, 17*, 43–64. Retrieved from:

- http://englishvls.hunnu.edu.cn/Downloads/LangTst/tst_003.pdf

Qichun, Z. (2005). The analysis of authenticity in language test. *CELEA Journal, 28*, 122–125. Retrieved from:

- http://www.celea.org.cn/teic/59/59-122.pdf

Central tendency

Central tendency (n.d.) Retrieved from:

- http://www.quickmba.com/stats/centralten/

Classical theory

Lin, C. J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. *The Journal of Technology, Learning, and Assessment, 6*. Retrieved from:

- http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1638/1473

Steyer, R. (n.d.). Classical (psychometric) test theory. Retrieved from:

- http://www.metheval.uni-jena.de/materialien/publikationen/ctt.pdf

Classroom observation

The value of observation (n.d.). Retrieved from:

- http://www2.education.ualberta.ca/staff/olenka.Bilash/best%20of%20bilash/observation.html

Classroom observation in teaching practice. In *Practice teaching: A reflective approach* (pp. 90–105). Retrieved from:

- http://www.professorjackrichards.com/wp-content/uploads/Practice-Teaching-A-Reflective-Approach-Chap-7-Classroom-Observation-in-Teaching-Practice.pdf

Classroom observation checklist

Classroom observation checklist (n.d.). Retrieved from:

- http://www.gadsdenstate.edu/faculty-and-staff/ie/documents/ClassroomObservationChecklist.pdf

Classroom observation instruments (n.d.). Retrieved from:

- http://www1.umn.edu/ohr/teachlearn/resources/peer/instruments/

Comparability

Interpreting claims of test comparability: Important considerations, concerns, and caveats (n.d.). Retrieved from:

- http://stepeiken.org/comparability

Dispersion

Dispersion – Deviation and variance (n.d.). Retrieved from:

- http://www.wyzant.com/help/math/statistics_and_probability/averages/dispersion_deviation_and_variance

Statistical dispersion (n.d.) Retrieved from:

- http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Statistical_dispersion.html

Effective testing and test utility fairness

Writing effective tests: A guide for teachers (n.d.). Retrieved from:

- http://www.glencoe.com/sec/teachingtoday/educationupclose.phtml/40

Fairness and bias. Retrieved from:

- http://www.siop.org/_Principles/pages31to34.pdf

Error

Error in measurement (n.d.). Retrieved from:

- http://www.regentsprep.org/Regents/math/ALGEBRA/AM3/LError.htm

Evaluation

Prakash, J. (n.d.) Valuable notes on the evaluation process in education. Retrieved from:

- http://www.preservearticles.com/201105216973/evaluation-process-in-education.html

Feedback, reporting, and results interpretation

Education and Staff Development, Queen Mary University of London (n.d.). Good practice guide on assessment and feedback to students. Retrieved from:
➢ http://www.learninginstitute.qmul.ac.uk/ideas/wp-content/uploads/Assessment-and-Feedback-Good-Practice-Guide.pdf

Guides to test reports (2012–13 edition) (2012). Retrieved from:
➢ http://www.celdt.org/resources/im/

Nicol, D. (2007). Principles of good assessment and feedback: Theory and practice. Keynote paper delivered at the Assessment Design for Learner Responsibility conference, 29–31 May 2007. Retrieved from:
➢ http://www.york.ac.uk/media/staffhome/learningandteaching/documents/keyfactors/Principles_of_good_assessment_and_feedback.pdf

Spiller, D. (2009). Assessment: Feedback to promote student learning. Hamilton, New Zealand: the University of Waikato. Retrieved from:
➢ http://www.waikato.ac.nz/tdu/pdf/booklets/6_AssessmentFeedback.pdf

Formative and summative assessment

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5–25. Available from:
➢ http://www.skolverket.se/polopoly_fs/1.126607!Menu/article/attachment/formative_assassement.pdf

Formative vs summative assessment (n.d.) Retrieved from:
➢ http://www.cmu.edu/teaching/assessment/howto/basics/formative-summative.html

Garrison, C., & Ehringhaus, M. (2007). Formative and summative assessments in the classroom. Retrieved from:
➢ http://www.amle.org/Publications/WebExclusive/Assessment/tabid/1120/Default.aspx

What is the difference between formative and summative assessment? (n.d.). Retrieved from:
➢ http://www.cmu.edu/teaching/assessment/basics/formative-summative.html

Frequencies

Frequency (n.d.). Retrieved 27 June 2013 from:
➢ http://en.wikipedia.org/wiki/Frequency

Hall, S. (n.d.) How to calculate frequency statistics. Retrieved from :
➢ http://www.ehow.com/how_7247964_calculate-frequency-statistics.html

Generalizability theory

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt & D.C. Howell, *Encyclopedia of statistics in behvioral science* (pp. 717–719). Chichester, UK: John Wiley & Sons. Retrieved from:
➢ http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/GTheoryEncyclo.pdf

Histogram

Histograms: Construction, analysis and understanding (2002). Retrieved from:
- http://quarknet.fnal.gov/toolkits/new/histograms.html

Data collection methods for program evaluation: Interviews (2009). Retrieved from:
- http://www.cdc.gov/healthyyouth/evaluation/pdf/brief17.pdf

Interview

Sewell, M. (n.d.). The use of qualitative interviews in evaluation. Retrieved from:
- http://ag.arizona.edu/sfcs/cyfernet/cyfar/Intervu5.htm

Qualitative methods (2011). Retrieved from:
- http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20190070~menuPK:412148~pagePK:148956~piPK:216618~theSitePK:384329,00.html

Data collection methods for program evaluation: Interviews (2009). Retrieved from:
- http://www.cdc.gov/healthyyouth/evaluation/pdf/brief17.pdf

Item response theory (IRT)

Item response theory (n.d.). Retrieved from :
- http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm

Item analyses, difficulty, and discrimination

Item analysis (n.d.). Retrieved from:
- http://www.washington.edu/oea/pdfs/resources/item_analysis.pdf

Item analysis (2003). Retrieved from:
- http://www.scrolla.hw.ac.uk/focus/ia.html

Item analysis of classroom tests: Aims and simplified procedures. Retrieved from:
- http://www.udel.edu/educ/gottfredson/451/unit9-guidance.htm

Item-centred and examinee-centred performance standards

Albano, T. D. (2008). The Angloff and Ebel standard setting methods.
Retrieved from:
- http://www.edmeasurement.net/5221/Angoff%20and%20Ebel%20SS%20%20-%20TDA.pdf

Eckes, T. (2012). Examinee-centred standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling,* 54, 257–283.
Retrieved from:
- https://www.testdaf.de/institution/pdf/publikationen/Eckes_PTAM_2012-3.pdf

Hazlett, C. B. (n.d.) Standard setting: Determining the appropriate Pass-Fail Score.
Retrieved from:
- http://www.oes.cuhk.edu.hk/Archives/Old%20Presentations/Standard.pdf

Näsström, G., & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment Research & Evaluation, 13*. Retrieved from:
➢ http://pareonline.net/pdf/v13n9.pdf

Item classification

Cognitive complexity classification of FCAT test items (2008). Retrieved from:
➢ http://fcat.fldoe.org/pdf/cog_complexity-fv31.pdf

Judgement

Brooks, V. (2012). Marking as judgment. *Research Papers in Education, 27*, 63–80. Retrieved from:
➢ http://wrap.warwick.ac.uk/2812/

Mark schemes, rubrics, and double marking

Grading and Performance Rubrics (n.d.). Retrieved from:
➢ http://www.cmu.edu/teaching/designteach/teach/rubrics.html

Swansea University (n.d.). Double Marking Policy. Retrieved from:
➢ http://www.swan.ac.uk/registry/academicguide/assessmentandprogress/doublemarkingpolicy/

University of Cambridge International Examinations (2008). Mark Scheme for the May/June 2008 Question Paper. Retrieved from:
➢ http://papers.xtremepapers.com/CIE/Cambridge%20International%20O%20Level/Bengali%20(3204)/3204_s08_ms_1.pdf

Moderation

New Zealand Ministry of Education (n.d.). Moderation. Retrieved from:
➢ http://assessment.tki.org.nz/Moderation

National Professional Standards for Teachers

Policy and Planning Wing, Ministry of Education, Government of Pakistan (2009). National Professional Standards for Teachers in Pakistan. Retrieved from:
➢ http://unesco.org.pk/education/teachereducation/files/National%20Professional%20Standards%20for%20Teachers.pdf

Normalization

Normalization (n.d.). Retrieved 27 June 2013 from:
➢ http://en.wikipedia.org/wiki/Normalization_(statistics)

Oral questioning

Improving your oral questioning techniques (2012). Retrieved from:
➢ http://www.protocol.co.uk/downloads/Improving_your_Oral_Questioning_Techniques.pdf

Peer assessment and self-assessment

Peer assessment (n.d.). Retrieved from:
➢ http://www.heacademy.ac.uk/hlst/resources/a-zdirectory/peer_assessment

Peer-assessment (n.d.). Retrieved from:
➢ http://www.cte.cornell.edu/teaching-ideas/assessing-student-learning/peer-assessment.html

Self-assessment (n.d.). Retrieved from:
➢ http://www.cte.cornell.edu/teaching-ideas/assessing-student-learning/self-assessment.html

Juwah, C. (2003) Using peer assessment to develop skills and capabilities. *USDLA Journal, 17.* Retrieved from:
➢ http://www.usdla.org/html/journal/JAN03_Issue/article04.html

Percentiles, stanines, percentile band scores, and grade equivalents

Percentiles (n.d.). Retrieved from:
➢ http://www.regentsprep.org/Regents/math/ALGEBRA/AD6/quartiles.htm

Robertson, I. (2004). Calculating percentiles. Retrieved from:
➢ http://www.stanford.edu/class/archive/anthsci/anthsci192/anthsci192.1064/handouts/calculating%20percentiles.pdf

Understanding stanines (n.d.). Retrieved from:
➢ http://www.nzcersupport.org.nz/marking/?p=75

Rogosa, D. (1999). *Accuracy of individual scores expressed in percentile ranks: Classical test theory calculations.* Retrieved from:
➢ http://cse.ucla.edu/products/reports/TECH509.pdf

Grade point equivalents (n.d.). Retrieved from:
➢ https://courses.cit.cornell.edu/thetr364/grading.html

Understanding grade equivalents and other standardized test scores (n.d.). Retrieved from:
➢ http://www.cobbk12.org/Lindley6/Uploads/Parents/UnderstandingGradeEquivalents.pdf

Placement tests

Placement test (n.d.). Retrieved 27 June 2013 from:
➢ http://en.wikipedia.org/wiki/Placement_testing

What are college placement tests? (n.d.). Retrieved from:
➢ https://bigfuture.collegeboard.org/find-colleges/academic-life/what-are-college-placement-tests

Performance standards

Anderson, J. (n.d.). *Accountability in education.* Paris: The International Institute for Educational Planning; Brussels: The International Academy of Education. Retrieved from:
http://www.iaoed.org/files/Edpol1.pdf

Projects and portfolios

Almanza, D., Cackley, P., Goins, K,. LeCloux, C., Lucas, K., Marston, T., … Vocelle, L. (1997). *Project based learning and assessment: A resource manual for teachers*. Arlington, VA: Arlington Education and Employment Program. Retrieved from:
- http://www.cal.org/caela/esl_resources/reepproj.pdf

Assessing project-based learning (n.d.). Retrieved from:
- http://serc.carleton.edu/introgeo/assessment/project.html

Portfolios for student growth (n.d.). Retrieved from:
- http://www.gallaudet.edu/clerc_center/information_and_resources/info_to_go/transition_to_adulthood/portfolios_for_student_growth.html

Prince George's County Public Schools (n.d.). Portfolio assessment. Retrieved from:
- http://www.pgcps.org/~elc/portfolio.html

Question setting

Assessment: Final assessment (n.d.). Retrieved from:
- http://www.cdtl.nus.edu.sg/handbook/assess/objective.htm

Question writing (n.d.). Retrieved from:
- http://examdesign.com/WriteQuestions.aspx

Rank order

Rank order scale (n.d.). Retrieved from:
- http://changingminds.org/explanations/research/measurement/rank_ordering.htm

Raw scores

Texas Education Agency (n.d.). Raw conversion tables. Retrieved from:
- http://www.tea.state.tx.us/index3.aspx?id=3270&menu_id=793

Reliability

Jacobs, L. (1991). Test reliability. Retrieved from:
- http://www.indiana.edu/~best/bweb3/test-reliability/

Phelan, C., & Wren, J. (2005–06). Exploring reliability in academic assessment. Retrieved from:
- http://www.uni.edu/chfasoa/reliabilityandvalidity.htm

Wells, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. Retrieved from:
- http://testing.wisc.edu/Reliability.pdf

Repeated observations (Spearman–Brown prophecy)

Brown, J. D. (2001). Can we use the Spearman-Brown prophecy formula to defend low reliability? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*, 7–11. Retrieved from:
- http://jalt.org/test/bro_9.htm

School improvement

Creemers, B. P. M., Stoll, L., Reezgit, G., & the ESI Team (n.d.). Effective school improvement – Ingredients for success: The results of an international comparative study of best practice case studies. Retrieved from:
➢ http://www.rug.nl/staff/b.p.m.creemers/effective_school_improvement_ingredients_for_success.pdf

Education Improvement Commission (2000). *School improvement planning: A handbook for principals, teachers, and school councils.* Retrieved from:
➢ http://www.edu.gov.on.ca/eng/document/reports/sihande.pdf

School improvement life cycle (n.d.). Retrieved from:
➢ http://www.advanc-ed.org/school-improvement-life-cycle

Standard scores

Standard score (n.d.). Retrieved from:
➢ https://statistics.laerd.com/statistical-guides/standard-score.php

Understanding standard scores (n.d.). Retrieved from:
➢ http://www.faculty.virginia.edu/PullenLab/WJIIIDRBModule/WJIIIDRBModule7.html

Standardization, standardized tests, norm referencing, and criterion referencing

Glutting's guide for norm-reference test score interpretation, using a sample psychological report (n.d.). Retrieved from:
➢ http://www.udel.edu/educ/gottfredson/451/Glutting-guide.htm

Zucker, S. (2003). Fundamentals of standardized testing. Retrieved from:
➢ http://www.pearsonassessments.com/NR/rdonlyres/20DB5E75-059E-4EB7-BFB1-F9C171ABE0C3/0/Fundamentals_of_Standardized_Testing_Final.pdf

Student progress

Safer, N., & Fleischman, S. (2005). Research matters/how student progress monitoring improves instruction. *Educational Leadership, 62*, 81–83. Retrieved from:
➢ http://www.ascd.org/publications/educational-leadership/feb05/vol62/num05/How-Student-Progress-Monitoring-Improves-Instruction.aspx

Taxonomy of objectives, Bloom's Taxonomy, and SOLO Taxonomy

Bloom's Taxonomy of educational objectives with verbs: Cognitive domain (n.d.). Retrieved from:
➢ http://www.madonna.edu/pdf/admissions/bloomcog.pdf

Bloom's Taxonomy of learning domains (1999). Retrieved from:
➢ http://www.nwlink.com/~donclark/hrd/bloom.html

Taxonomy of educational objectives (1956). Retrieved from:
➢ http://centeach.uiowa.edu/materials/Taxonomy%20of%20Education%20Objectives.pdf

Test construction

Devine, M., & Yaghlian, N. (n.d.) *Test construction manual: Construction of objective tests*. Retrieved from:

- http://www.cte.cornell.edu/documents/Test%20Construction%20Manual.pdf

Test construction (n.d.). Retrieved from:

- http://citl.indiana.edu/resources/teaching-resources1/teaching-handbook-items/test-construction.php

Wegener, D. P. (n.d.). Test construction. Retrieved from:

- http://www.delweg.com/dpwessay/tests.htm

Test specification

There are various resources on this website to explore:

- http://www.rewardinglearning.org.uk/qualifications/results.aspx?g=0&t=0&s=113&v=0&q=152&d=d

Topic coverage

Gomes, H. (2009). How much test coverage is enough? Retrieved from:

- http://www.summa-tech.com/blog/2009/03/24/how-much-test-coverage-is-enough/

Validation

Test validation (n.d.). Retrieved from:

- http://www.waldentesting.com/services/validation.htm

Validity

How do you determine if a test has validity, reliability, fairness, and legal defensibility? (2006). Retrieved from:

- http://www.proftesting.com/test_topics/pdfs/test_quality.pdf

Value added and school accountability

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Education Testing Service. Retrieved from:

- http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Loeb, S., & Figlio, D. (2011). School accountability. In E.A. Hanushek, S. Machin, & L. Woessmann (eds.), *Handbook of the economics of education, volume 3* (pp. 383–423). San Diego: North Holland. Retrieved from:

- http://cepa.stanford.edu/content/school-accountability

Murphy, D. (2012). Where is the value in value-added modeling? Retrieved from:

- http://educatoreffectiveness.pearsonassessments.com/downloads/ViVa_v1.pdf

Wei, H., Hembry, T., Murphy, D. L., & McBride, Y. (2012). Value-added models in the evaluation of teacher effectiveness: A comparison of models and outcomes. Retrieved from:
➢ http://www.pearsonassessments.com/hai/images/tmrs/ComparisonofValue-AddedModelsandOutcomes.pdf

The Florida Department of Education (USA) website provides examples of school accountability reports:
➢ http://schoolgrades.fldoe.org/

Variance and correlation

Bunzel, H. (2006). Variance, covariance and correlation. Retrieved from:
➢ http://www2.econ.iastate.edu/classes/econ500/bunzel/documents/mpdslides3.pdf

Relationship between correlation and sample variance (2012). Retrieved from:
➢ http://stats.stackexchange.com/questions/22890/relationship-between-correlation-and-sample-variance

z-score and T-score

Siegle, D. (n.d.). Standardized scores. Retrieved from:
➢ http://www.gifted.uconn.edu/siegle/research/normal/interpret%20raw%20scores.html

Higher Education Commission