IV INTERNATIONAL SUMMER SCHOOL
'Test Development in Education and Psychology: Theory and Practice'

PROGRAM & STRUCTURE

It is customary for our Summer Schools to start by an introductory lecture for all participants, where we briefly cover topics relevant for any assessment professional, and subsequently continue in two separate tracks.

Integrated part (common for both tracks): Introduction to Evidence-Centered Design (ECD) (Presenter – Mark Zelman)

ECD is an approach developed in ETS and is used to construct assessments in terms of evidentiary arguments. The lecture will give an overview of the evidence-centered design and its implementation. In particular, the lecture will cover the high-level ECD model: Conceptual Assessment Framework. In addition, the lecture will provide brief working definitions of certain technical concepts, in particular, proficiency model, evidence model, task model, and assembly model.

The language of this part is Russian (partial translation will be provided).

Track 1: Development of Assessment Instruments within the Evidence-Centered Design (ECD) Approach (Presenter – Mark Zelman)

The course covers the following topics:

Bayesian Statistics: The lecture will give a brief overview of how to develop and use Bayesian models for extending probabilistic reasoning about student's proficiencies beyond reasoning conducted in the traditional Frequentist statistical framework. The lecture will describe how Bayesian statistics can be a guiding principle for reasoning about complex psychometric models which are not tractable under the frequentist approach. In addition, the lecture will briefly touch some modern computing algorithms like Markov Chain Monte Carlo (MCMC) which can solve problems cast in the Bayesian statistical framework.

Graphical Models: The lecture will be concerned with answering a question of how the underlying principles of many assessment designs – the psychometric model should reflect the cognitive model in a manner that fits the purpose of the assessment – could be addressed by graphical networks. This lecture provides some basic foundations of graph theory and graphical models and links them to Bayesian networks (i.e., types of Bayesian networks – graphical models in which all of the variables are discrete – are very popular in artificial intelligence communities and gaining some popularity among psychometricians).

Putting It All Together: The lecture concerns with putting ECD, Bayesian statistics and graphical networks together for the purpose of describing evidence of student's KSA, task, assembly, and presentation models necessary for the assessment. The lecture will briefly discuss how to build Bayesian probability models that are embedded in a graphical structure that reflect our knowledge about the student's KSA, defined by ECD, and propagate evidence through the graphical model in order to make a claim about a student.

The language of Track 1 is Russian.

Track 2 will consist of two successive parts.

Track 2, Part 1: Evaluation of Test Quality (Presenter — Bas Hemker)

In order to function well, tests need to be of good quality. In this course many factors of the evaluation of tests are discussed. First, an overview of evaluation systems is given, showing their similarities and differences. Secondly, important indicators of test quality are reviewed that can be found in most, if not all, review systems. These include quality of the norms, reliability and validity. Finally we focus in more detail on topics related to the investigation and evaluation of test quality in practice.

The use of quality criteria of test is important is at least two ways. First they are suitable to guide test development in order to obtain the best possible measurement instruments. Secondly, they can be used as an instrument for internal and external audits.

There are a considerable number of tools to assess the quality of tests. They can be distinguished as guidelines, standards, reviews and evaluation systems. Special interest will be given to a number of systems that have a sizeable Dutch influence. Many of the tools to assess the quality of tests consider the same kinds to threats to the quality of the test or test use. In some cases these are presented as criteria on which the tests can be evaluated, such as in the Dutch COTAN system. In other cases the evaluation is structured around key issues such as the American AERA/APA/NCME (2014) Standards do (foundations, operations, applications). Also the quality of tests can be viewed completely from a validity standpoint.

In the last section of the course, many practical issues involving the practical evaluation of the test quality are discussed. For example, when are we evaluating the equivalence of paper and pencil tests with computer based tests, what research would that require? What kind of norms are to be considered relevant for what goals? How do evaluate DIF? When is DIF a problem and does the lack of DIF means there are no issues that can be related to DIF? What measures of reliability can we use, and when do we need to use what measure? What are relevant alternatives to single measures of reliability? How do we evaluate construct validity? What are validity arguments?

Track 2, Part 2: Constructing Performance-Based Assessments: Designing Task Specifications and Assessment Tools (Presenter – Carol Myford)

The goal of this training is to help participants become competent in devising and scoring performance-based assessments. These are assessments in which individuals demonstrate their abilities to use their knowledge and skills by creating a product, carrying out a process, or engaging in a performance. Examples of processes (or performances) might include giving oral presentations, carrying out a complex procedure, conducting an experiment, taking part in a debate, holding an interview, engaging in a dialogue, participating in an interactive simulation, repairing a piece of equipment, performing on an instrument or in a play, and so on. Examples of products might include lab reports, posters, videos, spreadsheets, term papers, audio recordings, drawings, models, brochures, business plans, and so on.

Participants will learn how to design the specifications for a performance-based task, discuss different types of performance-based assessments, their strengths and limitations. They will also identify and define the criteria that they want to use to evaluate performance. Participants will learn how to turn criteria into checklists and various types of rating scales (i.e., numerical, graphic, descriptive graphic) and rubrics (i.e., holistic, analytic, generic, task specific) that are useful for both formative (informal) and summative (formal) assessment purposes.

The language of Track 2 is English.

The tracks will include lectures, workshops, discussion groups, and individual consultations for students.