Center for Research on Learning and Instruction
UNIVERSITY OF SZEGED

H-6722 Szeged, Petőfi S. sgt. 30–34. Tel./ Fax: 36/62-544-354 e-mail: ok@edu.u-szeged.hu

SZEGED WORKSHOP ON EDUCATIONAL EVALUATION

Program

27–28 April 2009

Venue:
Szeged Committee of the Hungarian Academy of Sciences
H–6720 Szeged, Somogyi utca 7.

Program Chairs:
Benő Csapó and Detlev Leutner

SZEGEDI TUDOMÁNYEGYETEM
OKTATÁSELMÉLETI KUTATÓCSOPORT
www.edu.u-szeged.hu/ok

27 April 2009 – Monday

10^{00} Competence Diagnostics

Models of Competencies for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes

Johannes Hartig, University of Erfurt, Faculty of Education *and Andreas Frey,* Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN)

The Priority Research Program „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (“Models of Competencies for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes”) initiated by Prof. Dr. Klieme (DIPF) and Prof. Dr. Leutner (Duisburg-Essen University) was approved in April 2006 by the senate of the Deutsche Forschungsgemeinschaft (German Research Foundation). The duration of the program is six years. The program covers basic research from a cognitive psychological view as well as from the view of the different subject-related didactics. It includes the development of psychometric models and concrete technologies for the measurement of competencies. In the Priority Research Program, competencies are defined as context-specific cognitive dispositions for achievement that are functionally related to situations and requirements within certain domains. Competencies are acquired through experience and learning and can be affected through exterior interventions.

http://www.kompetenzdiagnostik.de/

Devising an online diagnostic assessment system

Benő Csapó, University of Szeged, Institute of Education

The University of Szeged Center for Research on Learning and Instruction has submitted a project proposal to develop an online diagnostic assessment system for the first six grades of the primary school. The purpose of this presentation is to introduce the project and to explore possible cooperation with other research groups.

Decomposing competences at students and classroom level by multilevel item response models

Johannes Hartig, University of Erfurt, Faculty of Education *and Jana Höhler,* DIPF

For the assessment of broadly defined competences, multidimensional item response theory (MIRT) provides an appropriate foundation to simultaneously model multiple basic abilities. If the assessed data is hierarchically structured, a decomposition of the latent (co-)variances in within- and between-group components is especially valuable for research questions focusing on persons embedded within a social system (e.g., family, class, or school) rather than on the single individual. Here, individual performance can be viewed as being affected, for example, by class-membership as a higher hierarchical level in addition to potential factors on the individual level. The main purpose of this contribution is to demonstrate the benefits of (co-)variance decomposition with an empirical application of a multilevel MIRT (ML-MIRT) model.

Data for this study are from a large-scale assessment of language awareness, reading and listening comprehension in English as a foreign language in the 9th grade (N = 10,059 students in 427 classes). The applied model is a multidimensional generalisation of the Rasch model, and contains three correlated dimensions on two

levels, one dimension for each construct. The first level is the individual level for students within classes, and the second level is the classroom level.
At the individual level a markedly more differentiated correlation structure is found compared with the classroom level. The calculated latent intraclass correlation coefficients indicate that more than half of the variance can be explained by differences between classes. Furthermore, this amount is reduced if school track is taken into account. For listening comprehension, the amount of explained variance indicates that it is more determined by classroom characteristics than reading comprehension and language awareness. Results and implications of this modelling approach for interpreting and communicating test scores, especially in the form of individual ability profiles for the assessment of a broader competence, will be discussed.

12^{30} LUNCH BREAK

13^{10} Assessing complex problem solving

German and international efforts

Joachim Funke, University of Heidleberg, Department of Psychology

The presentation gives a short introduction to complex problem solving and to different approaches of measurement for this construct. Besides some own ideas, approaches from other labs are used for illustration. At the end, the next theoretical, empirical, and methodological steps are outlined.

Computerized assessment of CPS

Samuel Greiff, University of Heidelberg, Department of Psychology

The growing interest in complex problem solving (CPS) increases the need for efficient assessment procedures. However, no psychometrically acceptable testing device is currently available and existing tests suffer from two major shortcomings: (1) little agreement on how to measure CPS on an individual level has been reached, and sound theoretical foundations cannot be found in existing tests; (2) tests consist of only one large scenario a participant has to work through. Thus, existing instruments contain exactly one excessive item and by that contradict basic psychometric requirements.
We introduce a new approach called MicroDYN that provides an infinite item pool based on a mathematical formalism. The type of items used requires participants to detect causal relations between two sets of variables and, subsequently, to control the system intuitively. Participants face several items each lasting about 5 minutes. By presenting minimally but sufficiently complex items, the problem of one-item-testing is overcome.
Additionally, MicroDYN is linked to Dörner's *Theory of Operational Intelligence*. Dörner (1986) names five abilities a test taker must meet to successfully solve complex problems: (1) reduction of information, (2) model building, (3) forecasting, (4) information retrieval, and (5) evaluation. Within the MicroDYN framework measures for three indicators (model building, forecasting & information retrieval) are empirically derived.
Psychometrical issues of MicroDYN are discussed. Its internal structure is evaluated by comparing competing models and data on the predictive power of external criteria is provided. The validity and the benefit of using several indicators simultaneously are shown empirically.

If – at least in the long run – complex problem solving can be nomothetically classified and established as a valid construct, it might be relevant in virtually all areas involving prediction or explanation of cognitive performance.

The Szeged assessments of CPS

Gyöngyvér Molnár, University of Szeged, Institute of Education

One of the recent trends in educational assessment is the shift of emphasis from curricular content to the measurement of cross-curricular competencies. Problem solving is such a competency; it has already been assessed in the framework of large-scale national and international projects as well. In the past years, we carried out several assessments to study the development of problem solving. This paper presents the results of these assessments. Altogether, ten age groups were assessed by the means of several problem solving tests and by the means of school-specific problems. Participants' age ranged from 9 (grade 3) to 17 years (grade 11). The different tests included a large number of common anchor items that allows the expression of all results on the same scales. The Rasch model was used for scaling the data and establishing a developmentally valid scale. Large differences were found at both individual and class level. These differences can mostly be attributed to the social background of students. The data also allows for the comparison of the results of students of the same age tested over time, and this comparison indicates a decrease in students' achievement. Comparing the results of disciplinary and CPS tests indicates the content-bound nature of students' knowledge and the difficulties of transfer. Achievement differences decrease within gender groups while they increase between the genders with age.

15^{00} COFFEE BREAK

15^{20} Assessing and Teaching 21st Century Skills

Assessing and Teaching 21st Century Skills – Project overview

Benő Csapó, University of Szeged, Institute of Education

Three ICT companies organized a task force and launched a global project to advance technology based assessment. This presentation introduces the entire project and prompts for cooperation.

http://www.atc21s.org/

Assessing and Teaching 21st Century Skills – Technological aspects

Jean-Paul Reeff, LIFE Research & Consult

The ACT21S initiative aims at identifying, teaching and measuring twenty-first-century skills in accordance with highest methodological standards. The presentation will tackle technological settings and requirements to implement this ambition at a system level, a classroom level and an individual level. Special emphasis will be given to differential requirements at the different levels and to the consequences for future action and cooperation.

28 April 2009 – Tuesday

10^{00} Preparations for an ESF project proposal

ESF project proposal on Technology-Based Assessment Systems in the 21st Century
Jean-Paul Reeff, LIFE Research & Consult

In 2006 several institutions submitted a proposal for an ESF-Network, to be led by DIPF, the University of Szeged and University of Luxembourg.
The proposed network aims at linking research and innovative development approaches from two different domains and their related sub-disciplines:

- Technology-based assessment research
- Research on distributed, collaborative learning environments of the future.

One major goal is to overcome traditional thinking in the design of technology-based assessments and to make full use of the latest technological developments in advanced learning environments, in order to shape more adequate instruments both for basic research and applied contexts. A second goal is to integrate advanced measurement theory and technology into emerging learning environments in order to improve a scientifically based measurement of learning progress. Despite good to excellent reviews the proposal was turned down. The presentation aims at outlining a relaunch of the network proposal, maybe in the context or related to the ATC21S initiative.

12^{30} LUNCH BREAK

13^{10} Experiences with the implementation of TAO

Recent progress – overview
Patrick Plichart, Centre de Recherche Public Henri Tudor

The definition and modeling of specific assessment processes is an essential condition for success in large-scale studies and national school system monitoring. In order to achieve a satisfactory level of quality and comparability of tests and assessments, several authors and practitioners have stressed the urgent need of standards and validated metrics applicable all along the course of this complex multi-step process. This diversity and the lack of common universal agreement arise from the variety of assessments targets and contexts. Reaching such a level of agreement is obviously largely unfeasible and no predefined system (process, models, and metrics) can reasonably be foreseen. As a consequence, the major challenge is rather to provide a framework of tools, methods and quality meta-models enabling the quality assessors to create and deploy their own processes and models according to their very specific objective, scope and context of assessment. The TAO-Qual project aims at providing a computerized framework implemented into the TAO platform, enabling assessment managers to design their assessment processes. It also aims at providing the workflow engine driving those processes giving the right task and tool to the right person at the right moment. The test as taken by subjects is in itself a complex process involving a specific sequence of activities, making use sometimes of decision-based branching, complex data inference and consistency checks. Therefore, the TAO-Qual workflow

engine may be used as well to design complex tests as processes. This presentation describes new features implemented into the TAO platform with respect to process-oriented design of assessment.

The DIPF experience
Jean-Paul Reeff, LIFE Research & Consult

The DIPF TBA project had originally been conceived to support the DFG priority programme on competencies assessment. After discussions with the German research ministry its size and scope has been substantially increased to impact on national and international programmes far beyond the priority programme. The presentation will review experiences over the last two years and lessons (to be) learned. After a strong focus on large-scale assessment in a first phase, the DIPF TBA strategy will face new challenges, among them the CIM initiative and related activities. The presentation will outline some of these new challenges, above all in relation with the planned Hungarian project.

The Szeged experience
Krisztina R. Tóth, University of Szeged, Graduate School of Educational Sciences

The presentation provides an overview on the experience with the adaptation/use of the Testing Assisté per Ordinateur (TAO) platform in the testing period of 2008. The aim of the project was to identify achievement differences between paper-and-pencil and computer-based test results, and to identify factors which could affect test results on different test media, or which may be responsible for differential item functioning. The sample consisted of fifth graders. Students were asked to solve the same inductive reasoning test created by Benő Csapó on both testing media. In the second half of the presentation I will outline the plans for the next testing period, May and June 2009. In this new measurement second and sixth graders students are involved, their skills and knowledge is going to be measured in both an online and a traditional environment with differential research designs. The object of the new project is to compare results across media and determine factors which influence students' achievement in computerized testing as regards different measured areas, samples, item types and research designs.

15^{00} COFFEE BREAK

15^{20} TBA in international assessment projects – PISA, PIAAC

The impact of CBA on item development process
Marilyn Binkley, University of Luxembourg

CBA Item Builder - a new Generation of Test Item Authoring Tools in Computer Based Assessment
Ingo Barkow, DIPF *and Michel Dorochevsky,* SOFTCON IT-Service GmbH

We present a new generation tool for building test items using an authoring tool which can be handled without any programming knowledge. We report on our experiences in the PISA and PIAAC projects, demonstrate the features of the CBA Item Builder using examples and summarize the benefits. We close with an overview of the architecture and the components involved in the process, from design up to delivery.

Data Warehouse functionalities in the NEPS study: Introduction of the database structure and EduDDI as a standard for meta data

Ingo Barkow, DIPF

Though the National Education Panel Study (NEPS), a German national longitudinal survey regarding all age groups with a projected timeline of 22 years, will be performed by paper & pencil only, implementing the database can be considered a challenge. To support this projected running time the XML-based EduDDI standard will be used for storing meta-data. Furthermore, a Datamart or Data Warehouse will be built for analyzing results of the survey. This presentation will show in short manner key features of EduDDI and the NEPS Data Warehouse.

National school monitoring and research on CBA based on TAO

Romain Martin, University of Luxembourg

The presentation will outline the use of the TAO platform in the context of the Luxemburg school monitoring. It will also present ongoing research projects that are based on the TAO platform and that try to address the question of the design of more innovative item formats that take greater advantage of the multimedia and interactive possibilities that are offered in the context of computer-based testing. The potential added value of innovative item types and the methodological challenges that are associated with it will be discussed.

Further themes

A comparative analysis of academic mathematicians' conceptions and professional use of technologies in university mathematics

Zsolt Lavicza, Faculty of Education University of Cambridge

There are diverse beliefs and assumptions about how and how much mathematicians use technologies to teach mathematics at the university level. However, as opposed to the elementary and secondary education, where large-scale studies regularly assess the extent of technology use, little is known about the integration of technology in university-level mathematics teaching and learning. In this talk, drawing on my dissertation research, I aim to outline some current practices of mathematicians' use of technologies, particularly Computer Algebra Systems (CAS), their views about the role of CAS in future mathematics teaching and students' mathematical literacy, and some factors that influence mathematicians to integrate technology into their own teaching practices. This research is based on interviews and questionnaires of more than a thousand mathematicians in Hungary, the United States, and the United Kingdom.

Multidimensional Adaptive Testing

Andreas Frey, Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN)

In many countries, the results of large-scale assessments like PISA, TIMSS, or PIRLS have recently received a lot of attention and have stimulated intensive and productive discussions about the effectiveness of educational systems. However, the valuable results of these studies come at a rather hefty price, since large sample sizes of typically thousands of students within each country have to be tested with tests which take several hours to complete. To ensure the cooperation of schools and students in the long run, and to limit costs, ways to increase the efficiency of the testing procedures used while maintaining the high level of precision and interpretability of the results should be examined. One possibility to foster measurement efficiency without losing measurement precision lies in computerized adaptive testing (CAT).

CAT is a special approach to the assessment of latent abilities in which the selection of the test items that are presented to the examinee is based on the responses the examinee gave to previously administered items. The aim of this selection procedure is to tailor the item presentation to the ability of the examinee, thus increasing the amount of information that can be extracted from the item responses. Compared to a conventional test with a fixed number of items in a fixed order (fixed item test, FIT), the number of items can typically be reduced by about one half without a loss in measurement precision when CAT is used. The gains in measurement efficiency can be further enhanced if the traditional one-dimensional approach of CAT is expanded to multidimensional adaptive testing (MAT).

The talk presents the results of an extensive simulation study examining the measurement efficiency of MAT compared to CAT and FIT for typical large-scale assessment situations.

Under the condition of five dimensions, correlated with each other by .85, the measurement efficiency (calculated by the ratio of the inverse of the mean squared error and the number of items used) for MAT was 1.3 times higher than for CAT and 3.7 times higher compared to a random item selection. Thus, measurement efficiency can be increased substantially by MAT compared to FIT and CAT. Practically speaking, the mean number of items presented to the examinee can be reduced by 73% without losing measurement precision. Thus, MAT represents a promising candidate for substantially reducing the testing effort associated with large-scale assessments.

Based on the results from the simulation study, practical considerations for using MAT within large–scale assessments are discussed.

ORGANIZERS AND HOST INSTITUTIONS:

Institute of Education
Graduate School of Educational Sciences
Hungarian Academy of Sciences