MAGYARORSZÁG MEGÚJUL

A projektek az Európai Unió támogatásával valósulnak meg.

UNIVERSITY OF SZEGED
CENTER FOR RESEARH ON LEARNING AND INSTRUCTION
www.edu.u-szeged.hu/ok

TÁMOP 3.1.9-08/1-2009-0001
„Developing Diagnostic Assessments” project

4TH SZEGED WORKSHOP ON EDUCATIONAL EVALUATION

ABSTRACTS

24-25 April 2012

Venue:

Szeged Committee of the Hungarian Academy of Sciences

7, Somogyi street, Szeged

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

SZEGEDI AKADÉMIAI BIZOTTSÁG
SZAB•1961

ABSTRACTS

24th April 2012

Session A

Eckhard Klieme	Feedback to Students: Experimental Studies on Assessment-Based Interventions

Feedback is known to be one of the most effective interventions in education. There is also quite some knowledge about criteria for good, effective feedback. E.g., feedback should be informative rather than controlling, criterion-referenced or ipsative rather than based on social comparison, refer to specific problem solving processes and outcomes rather than personal traits, and it should provide advice on how to improve. However, most of this knowledge is based on psychological experiments in the laboratory which oftentimes use tasks that are quite different from classroom-based learning. So, further research is needed to find out how feedback can be embedded in education, and what kind of assessment will provide an optimal foundation for effective feedback.

These research questions were addressed in a project called "Conditions and Consequences of Classroom Assessment" (Co²CA). Up to now, the project saw four stages: (1) Mathematics items were tried out and scaled in a large survey in lower secondary schools. On top, students and teachers were asked about assessment and feedback practices. Results show that feedback mainly based on grades is negatively linked to student motivation and achievement. If a student perceives feedback as ipsative, both motivation and achievement tend to be stronger. (2) Students who had participated in the survey were randomly assigned to different kinds of feedback. As expected, criterion-referenced feedback had a positive effect on attributions and student satisfaction. (3) The assessment material and feedback procedures developed in the large survey were later used in a laboratory experiment. Here, a process-oriented type of feedback had an indirect (mediated) positive impact on motivation and achievement. Also, assessments which had a close connection to content being taught recently were perceived as more helpful than broader, standards-based assessments. (4) Finally, a longitudinal study in 40 classrooms aimed at ecologically valid effects of feedback practices on student achievement and motivation. In fact, feedback practices reported by students explained a significant part of cognitive and motivational change during a unit of mathematics teaching and learning. However, an experimental intervention based on process-oriented feedback provided by the teacher had no significant effect.

These findings are discussed in terms of (a) rules for effective use of formative assessment in schools, (b) problems in designing interventions based on teacher training. Although we can prove that classroom practices in assessment and feedback do have measurable effects, it seems to be very difficult to change these practices in a sustainable way, strong enough to identify significant effects on student learning.

Benő Csapó	Second Phase of the Implementation of an Online Diagnostic Assessment System

The Center for Research on Learning and Instruction at the University of Szeged started to implement the second phase of the "Developing Diagnostic Assessment" project. This phase is planned to be carried out in 2012-2014, and the activities of this phase, alike to the first are organized into seven work packages. The first WP continues the framework development in reading, mathematics and science. The second WP develops online assessments at further domains, seven domains (visual skills, social skills, English as foreign language, civic education, motivation, learning to learn, health literacy) were already explored in the previous phase, and seven new ones (writing, ICT literacy, financial literacy, musical abilities, problem solving, combinative reasoning, inductive reasoning and creativity) will be added in this phase. WP 3 deals with the development of the online platform and building item banks. We begin to connect intervention to assessment and develop materials, mostly online games to develop the skills the assessment indicate as weak, and this is the task of WP 4. In-service training of teachers (specialists, e.g. item writers, experts and users of the system in the practice) remains a massive part of the project (WP 5). Further work packages (WP 6) will focus on developing feedback systems for students and teachers, and with developing policy frameworks to support the implementation at school and at school district level. This presentation introduces the planned work and overviews the work packages.

Session B

Samuel Greiff and Joachim Funke	Interactive Problem Solving and its Realization in Large-Scale Assessments
In PISA 2012, complex problem solving will receive special attention for at least three reasons: (a) It is an additional option for the participating countries (46 out of 68 voted for it), (b) it is measured for the first time computer-based, and (c) problem solving is understood as Interactive Problem Solving and uses dynamic systems. In our presentation, we focus on the dynamic systems approach for competence assessment. Interactive problem solving requires from participants to explore and control minimal but sufficient complex systems like remote control, mobile phone, or home appliances. To model such systems, we use the formalisms of structural equation systems and finite state automata. With both approaches, a psychometrically sound assessment of the three theoretically derived facets “information retrieval”, “model building”, and “forecasting” is possible and useful for the description of interactive problem solving.	
Sascha Wüstenberg Samuel Greiff, Joachim Funke, Gyöngyvér Molnár, Andreas Fischer and Benő Csapó	Measurement Invariance of Complex Problem Solving Ability Measured by MicroDYN
Complex Problem Solving (CPS) is seen as a cross-curricular competency, which has recently attracted interest in large-scale-assessments. In the PISA 2012 cycle, CPS will be assessed across the world partly using minimally complex computer simulations based on the MicroDYN approach. In this talk, we present empirical results on MicroDYN, a computer-based test containing multiple items aimed to measure two main complex problem solving processes - knowledge acquisition and knowledge application. More specifically, we (a) test a measurement model for MicroDYN composed of the two aforementioned processes, we (b) evaluate whether measurement invariance holds across different grade levels of high school students in order to compare latent means between grade levels and (c) investigate relations between CPS, reasoning, grade point average (GPA), and parental education. Analyses are based on N=855 Hungarian high school students in grades 5 to 11. Using structural equation models, results show that (a) a 2-dimensional model with the facets knowledge acquisition and knowledge application fit the data best. Furthermore, (b) CPS is strongly factorial invariant and thus, mean differences between grades can be interpreted meaningfully. Latent means in both facets increase at higher grades, disregarding a considerable drop in CPS performance in grade 9. Finally, (c) results based on path analyses indicate that knowledge acquisition incrementally predicts variance in parental education and – to a lesser extent – in GPA even beyond reasoning. A proper assessment of CPS in addition to subject related abilities is of high educational relevance. Results of this study provide important implications on how to use CPS in an assessment context.	

Session C

Andreas Fischer, Sascha Wüstenberg and Samuel Greiff	Measuring Complex Problem Solving as a Five Dimensional Construct

Complex Problem Solving (CPS) has a tradition of experimental research lasting almost four decades. Complex problems feature a set of characteristic demands: (a) the polytely of the task at hand, (b) the complexity of its structure, (c) the interconnectivity of the involved variables, (d) the intransparency of these connections, and (e) the dynamics of the system.

Each of these features aims at a corresponding facet of an individuals' complex problem solving competency: (a') Evaluation and priority setting in case of conflicting goals (= evaluation). (b') Focusing on the most relevant information (= reduction of information). (c') Successive generation of an adequate internal representation of the system structure (= model building). (d') Active exploration of important information (= information retrieval). (e') System control in light of system dynamics and past decisions (= system control).

The computer-based test MicroDYN allows for assessing three of these five facets of CPS competency („information retrieval", „model buidling" and „system control").

In a recent study we integrated indicators for the two missing facets („reduction of information" and „evaluation") in MicroDYN: N=92 university students completed this version of MicroDYN and first results support the psychometric quality of the new indicators regarding Cronbach's α (.774 < α < .899) and empirical separability of the facets (r < .632 after correction for attenuation). Implications will be discussed.

Ursula Pöll, Julia Hilse, Jonas Müller and Samuel Greiff	*An Authoring Tool to Create Complex Problem Solving Items: the Item-Builder*

In 2009 the Leibniz Institute for Educational Research and Educational Information (DIPF) in Germany developed in cooperation with SOFTCON – a IT-service provider stationed in Munich – a software, in which MicroFIN and MicroDYN items can be implemented to assess Complex Problem Solving: The CBA Item-Builder. By means of this program conceptual ideas pertaining to Complex Problem Solving can be transformed into a computer based measurement device in an automated and easy-to-use way. The Item Builder was especially developed to provide test developers with a maximum amount of freedom when developing items. It can be operated without any programming knowledge.

The implementation of a basic item will be demonstrated, in order to exemplify some og the features within the Item Builder such as defining states and variables, integrating graphics, buttons and texts and linking events with states and variables.

Julia Hilse, Jonas Müller, Ursula Pöll and Samuel Greiff	*Project Introduction: Interactive Problem Solving and Lifelong Learning*

Recent developments in the assessment of interactive problem solving (IPS; synonym to Complex Problem Solving) have yielded considerable returns in terms of reliable and efficient measurement devices and theoretical clarifications (cf. Greiff, 2012, Greiff & Fischer, submitted, Greiff & Funke, 2009, Wüstenberg Greiff & Funke, 2012). As a result there is growing interest outside the classical realm of educational assessment in using the measurement tools of IPS as a marker for general learning activities e.g. in the area of lifelong learning (Ederer, Warnke, Greiff & Schuller, in press).

The LLLight'in'europe project (www.lllightineurope.com) seeks to utilize these advancements to clarify connections between corporate strategies concerning lifelong learning, public policy environments promoting these activities and measures of success as innovations, economic returns and skill acquisition. Involving international partners from the fields of economics, management, education and psychology about 60 companies from different countries and different cultural backgrounds constituting a total of over 4000 individuals as well as public policy officials representing 15 institutional environments will be assessed in relation to their successful interaction with complex and dynamic environments hence using IPS skills as a marker for lifelong learning.

In addition to a more profound understanding of IPS and its relation to various highly relevant indicators of economic and social success, the study will establish reference values of IPS skills covering a large spectrum of countries, cultures, companies, hierarchical positions and professions.

Session D

Daniel V. Holt and Joachim Funke	*Using the Tailorship MIcroworld Simulation to Measure Complex Problem Solving Ability*
The Tailorshop simulation is a computer-based decision-making and problem solving task, in which participants have to manage a virtual shirt factory. It was one of the earliest microworld simulations developed to study complex problem solving and has since been used in a range of experimental and psychometric studies. The general plausibility and arguably high content validity of the scenario still render it an interesting option for current and future studies of problem solving. We will first introduce and briefly demonstrate the Tailorshop simulation and then summarize existing research pertaining to its psychometric characteristics. In particular, we will focus on recent empirical findings addressing the psychometric reliability of the task, its convergent validity in relation to other measures of problem solving, and its criterion validity with respect to several external criteria. We will conclude by discussing the suitability of the Tailorshop simulation for different research and assessment requirements and how this task may inform the development of future measures of complex problem solving ability.	
Andreas Spredemann and Samuel Greiff	*Assessing Developmental Aspects of Combining Domain-Specific and Domain-General Problem Solving Ability: Some Ideas*
Until now, the MicroDYN approach, a measurement device for Complex Problem Solving, was used for the assessment of general problem solving ability. However, items within the MicroDYN approach have the potential of embedding items into a specific contexts involving a higher amount of prior knowledge into the solution process and thereby addressing additionally domain specific aspects of problem solving. In our presentation we like to present ideas on how the semantic embedment could affect different processes of problem solving and what could be the benefits of a combined measurement of domain specific and general problem solving. One domain in which such items are easily realized is natural science, more specifically physics, chemistry and biology. Within this context it is conceivable that previous knowledge has an impact on the effect of different domain specific problem solving processes. Domain specific previous knowledge could influence the effect of general problem solving ability on domain specific performance. That is, previous knowledge enhances overall performance and, simultaneously decreases the impact of general processes. Assessing this effect within a coherent assessment framework will yield important insights into the roles of domain specific and general cognitive processes. First ideas are presented in this talk.	

Session E

Sirkku Kupiainen and Mari-Pauliina Vainikainen	*Effort, Interest and Perceived Attainment in Low-Stakes Assessment. Differences Between Fourth, Sixth and Ninth Graders as Test-Takers*

Since the 1990s, there has been a growing interest in assessing not just curricular achievement but the more general cognitive and affective goals of education. The most well known endeavour is the triennial OECD PISA study, but also the European Union project for defining key competencies for lifelong learning – including learning to learn, the framework of the present study – and the 21st century skills assessment project are expanding the notion of what is seen as salient outcomes of education.

Common to all these projects at the level of assessment is their mixing of cognitive tasks and of self-report questionnaires pertaining to diverse factors believed to indicate readiness for new learning and successful adaptation to the rapidly changing demands of the future. Additionally, due to their non-curricular nature, the studies are generally 'low-stakes' at student level, meaning that there are no ramifications for the assessed students based on their performance. Accordingly, the quality of the data is dependent on students' willingness to 'lend themselves' for its production.

In the Finnish framework, learning to learn refers to 'the diverse cognitive and affective factors that are central to the application of existing skills to novel tasks and to new learning' (Hautamäki et al., 2002, p. 5), and the instrument developed for its measuring comprises cognitive tasks and self-report questionnaire scales pertaining to diverse affective factors allegedly contributing to new learning (Hautamäki et. al, 2002, 2006, 2010; Kupiainen et al., 2011). Building on the experience of a European learning to learn pilot project in 2008 (Hoskins and Fredriksson, 2008; Kupiainen et. al., 2008), two sets of questions addressing some aspects of metacognitive monitoring and accuracy (interest, effort, and perception of one's performance) have been included in the latest version of the Finnish test (Moreno, 2002).

Aim of the study

The current presentation will focus on students' views regarding the interest and their own performance in the cognitive tasks in view of their more general school related attitudes (expressly agency: effort and academic self-concept), both in relation to their performance in the respective tasks. Furthermore, possible group differences in and development of students' task-specific attitudes will be looked at.

For the fourth grade cohort, part of an ongoing longitudinal study began at grade one, also earlier data with e.g. standardised test results at grade 2, will be used.

The underlying interest of the presentation is to bring fourth into discussion the role of contextual affective factors (interest, willingness to apply effort) regarding students' performance in a low stakes test in which the tasks are related to but distanced from everyday school tasks.

Data and methodology

The data is from two separate large scale studies in the Helsinki metropolitan region, collected in 2010 (2000 students for the 6th and 9th grade cohorts, 1000 for the 4th). The assessments were implemented at regular lesson time, with mandatory participation even if the teacher did not explicitly control students' work. In the majority of schools, the 90 minute assessment was administered at the same time in all classes and was overseen by the classroom teacher (4th and 6th grade) or another teacher familiar to the class (9th grade).

Structural equation modelling will be used to look at the respective role of the task-specific and more general metacognitive/affective factors in explaining students' performance.

Results

The studies are under work at the moment so results will only be presented in the workshop.

Edit Tóth and Benő Csapó	*Teachers' Attitudes Towards Assessment and Accountability*

In the economically developed countries the evaluation of students' achievement with standardized tests in international and national context has an essential role in monitoring public education, in policymakers' decision-making and in comparing educational systems or institutions (OECD, 2007). Depending on the purpose of the survey certain stakes are connected to the large-scale achievement tests. Increasing number of countries use mandatory assessments, which form the bases of accountability. Due to the stakes, achievement assessments may have a lot of positive and negative effects on teaching and learning process (e.g.: Koretz, 2005; Stecher, 2002, Hanushek, 2005).

In our study we analyze elementary (N=1212) and secondary (1147) school teachers' opinions on students' performance-measurements on the basis of the results of a national representative questionnaire-survey. After examining teachers' attitudes on the usefulness of national and international measurements on a system-level, we focus onto mapping out teachers' view on some aspects of the National Assessment System, since it serves as a basis of the Hungarian performance-based accountability system as it provides the stakeholders involved in education with feedback on the efficiency of the work carried out in the institutions on the basis of students' test scores.

Our results show that teachers find the system-level measurements important, but they have doubts about the generalizability of the conclusions that can be drawn from the results. Teachers feel mostly worried because of the National Assessment System out of the measurements carried out at different levels. Primary school teachers consider preparation significantly more important than secondary school teachers. They have changed many aspects of their practice: look for more efficient educational methods, pay more attention to requirements regarding their professional development and developing the components of the measured competencies. The results point out that teachers accept the system-level students' performance-measurements as these entities have already penetrated our pedagogical culture. At the same time, the most important task is to facilitate teachers' work with the feedback gained from the measurements.

Session F

Ingo Barkow	Metadata Models for Process Mining and First-Levels Analysis of Studies
When researchers conduct their studies they generate metadata in every step of preparation and processing. The range goes from metadata about study information, study design, building of instruments, data collection, dissemination, down to archiving. All of these processing steps generate valuable metadata information for secondary usage. Unfortunately most of these information currently gets lost as it is documented neither in the codebook nor the final dataset. Recently the introduction of research data facilities leads to a tendency towards standardization and recording of metadata (e.g. DDI, SDMX, QTI, APIP). Nevertheless there are still 'white spaces' especially in the field of educational sciences and computer-based platforms. This talk will introduce some considerations about two of these currently underdeveloped topics - process mining and first-level analysis of studies. In process mining the output of log files from computer-based assessment tools is analyzed to find certain patterns (e.g. to predict the outcome of an item by timing information). Metadata about process mining describe a model how a structured log file and surrounding information should look like to enhance these analyses. The other consideration is to develop a model how first-level analysis of the dataset can be described by a research team (normally the original conductor of the study) in a meaningful manner so another research team doing secondary analysis can follow up on the procedures (e.g. Stata, SPSS or R scripts) which originally generated certain variables. A mechanism to document original research processes would be a gain for every research data facility.	
Krisztina R. Tóth, Heiko Rölke and Frank Goldhammer	Clustering Students on the Basis of Test-Taking Behaviour in Internet-Based Simulations
Current educational assessments increasingly aim at getting valid and reliable data derived from computer-based online simulations (e.g. OECD-PISA Electronic reading assessment, or OECD-PIAAC problem solving in technology-rich environment). These electronic resources intend to face students with everyday problems like ordering a book or a flight, writing/sorting emails, etc. Due to the complexity of test materials, students can interact with tasks and included stimuli in various ways and multiple steps. In order to be able to reproduce and further analyse individual interactions, students' activities are stored in automatically generated log files or databases. The present investigation is based on log data from a study of Pfaff and Goldhammer (2011) that assessed students' skills in evaluating information retrieved from the World Wide Web with respect to credibility. The logged test-taking behaviour in these Internet simulations serves to demonstrate how to use data on the test-taking process for evaluating students' interactions with simulation-based assessments. Furthermore, we give an example of clustering students on the basis of test-taking behaviour to identify similar behaving groups of secondary school students. Our examination bases on complex simulation items, which implement web search engines and enable us to measure skills required in accessing and evaluating information on the Internet. First results of our empirical investigation are presented.	

Session G

David Tobinski	From Problem Space to Planning Space - the Tower-of-Hanoi reloaded
The assessment of complex problem solving (CPS) in the adolescence is a challenge for cognitive psychology. Regarding the fact, that previous research are focused on adulthood and hardly used systems with eigendynamics, the paradigm EcoSphere (ESP) has been developed. The ESP technology uses a system of differential equations. The exploration of the system takes place in a vivid simulation and generates a new quality of ecological validity. This fact qualifies the assessment of CPS, to bring up scenarios with a semantic embedment near the curriculum of the participants. The substantial progress in creating curriculum-near scenarios lies in the possibility to analyze an additional and central factor of complex problem solving: the previous knowledge. Regarding the importance of knowledge the ESP technology uses a new highly standardized computer-based test for causal diagrams. The ESP technology brings up four different designs: (1) instruction of a complex system; (2) observing a complex system; (3) exploring a complex system and (4) controlling a complex system. The control performance and the construction of mental models will be tested with different Items with de- and increasing complexity.	
Gyöngyvér Molnár and Attila Pásztor	Game-Based Development of Thinking Skills
Computer games should have a significant role and offers new opportunities in educational evaluation. Development of abilities and reasoning skills via computer games is promising in many domains. The purpose of this pilot study is twofold: to investigate the opportunities and effectiveness of applying educational games to improve students' reasoning skills, second compare the possibilities of the analyses in face-to-face and game-based environment. The experimental group constituted of 123 first- and second-grade students, whereas the control group had 137 students having similar background variables. One third (38 students) of the experimental group took part in a game-based training in computer-based environment, while the remaining part participated in a face-to-face training. The training program consists of 120 educational games designed for young children. In the face-to-face training there was a team of implementers, namely the class teachers who introduced the activities to the children, while in the computer-based training the games were presented via touch screen monitors using headset in classroom setting. The effectiveness of the training was measured with an inductive reasoning test, comprising 37 figural, non-verbal items (Cronbach α=.87). Besides the test-based data collection innovative assessment technologies were used in the game-based environment with the intent of monitoring the affective processes, e.g. logging and analyzing metadata, such as head movement and facial expressions. There was no significant change in performance between pre- and post-test in the control group (t=1.44, p>.05), while the experimental group managed to achieve significant development in the experimental period (t=-18.8, p<.00), significantly outperforming the control group by more than one standard deviation on the post-test. There were no significant differences between the post-test mean scores of students getting the training in face-to-face or game-based environment (t=1.70 p>.05). According to the student level analyses there were no students in the experimental group whose performance dropped significantly from pre-test to post-test; moreover, several students improved in both modalities by more than one standard deviation. The effect size of the training program in face-to-face environment was d=1.05, in game-based environment d=.87 (both p<.01). The most frequent facial expressions were surprise (31%), happiness (24%), and anger (16%) during game-based training. Disgust (7%), fear (3%), and sadness (1%) were less frequent. The distributions of the facial expressions did not show significant relationships with the developmental level of reasoning skills.	