UiO : CEMO – Centre for Educational Measurement
University of Oslo

Presents

Frontiers in Educational Measurement

September 12 and 13, 2018

Conference Abstracts

Keynotes

From adaptive testing to adaptive learning

Hua-Hua Chang

Wednesday, September 12, 2018; 9:00 - 10:00, Room: FORUM

Modern theories in educational assessment are rapidly transforming testing from unaccommodating ranking measures into flexible and informative tools that can be used to address the compelling needs of various stakeholders in education. The presentation will start with a historical review of some theoretical developments of Computerized Adaptive Testing (CAT). Then, we will discuss how the cutting-edge testing technology can facilitate individualized learning. Our focus will be on how to build a reliable, and also affordable, adaptive tool for schools to classify students' mastery levels for any given set of cognitive skills that students need to succeed. Results from some experiments concerning the potential of using the CAT technology for both diagnostic and summative purposes will be presented.

Current challenges for the modelling of educational data

Harvey Goldstein

Wednesday, September 12, 2018; 14:30 - 15:30, Room: FORUM

Traditional ways of acquiring data to answer research questions are being challenged in a number of ways. The increasing availability of very large population based administrative datasets provides new and exciting possibilities for more detailed and extensive analyses that can handle the real life complexity of the data and with sufficient power to detect interesting interactions with consequences for increasingly informative interpretations. A major challenge for data methodologists is thus to develop and make widely available statistical models that are able to reflect this complexity. At the same time, studies that are designed to collect more nuanced information than that found in administrative datasets, remain important but are increasingly subject to ‘non-response’ which can lead to potential bias and threats to validity. This too poses a serious challenge to data analysts. The talk will discuss these two trends, their interrelationship and some of the emerging methodologies needed to address the issues. Specifically, the talk will look at new ways for modelling data with missing values, ways for incorporating knowledge about unreliability within a model, issues in the linking together of data from different sources such as education and health, and implications for the ways in which research studies, especially longitudinal ones, are designed. Illustrations of data analyses will be given.

The argument for a “Data Cube” for large-scale psychometric data

Alina von Davier

Thursday, September 13, 2018; 8:30 - 9:30, Room: FORUM

In the recent years, the work with educational testing data changed due to the affordances provided by the technology, the availability of large data sets, and by advances made in data mining and machine learning. Consequently, the data analysis moved from traditional psychometrics to advanced psychometrics to computational psychometrics. In the computational psychometric framework, the psychometric theory is blended with the data-driven knowledge discovery. Despite the advances in the methodology and the availability of the large data sets collected at each administration, the way the data (from multiple tests at multiple times) are collected, stored and analyzed by the testing organizations is not conducive to these real-time, data intensive computational psychometrics and analytics methods that can reveal new patterns and information about the students.

In this presentation, I am proposing a new way to label, collect, and store data from large scale educational learning and assessment systems (LAS) using the concept of the “data cube” introduced by data scientists about 10 years ago to deal with stratification problems in big data in marketing contexts. However, applying the concept to the educational data is quite challenging: The challenges are due to the lack of coherence of the traditional content tagging, of an identity management across testing instruments, of collaboration between the psychometricians and data scientists, and most recently, the lack of validity of the newly proposed machine learning methods for measurement. Currently data for psychometrics is stored and analyzed as a two-dimensional matrix – item by examinee. The items’ content, the standards or taxonomies are usually stored as narratives in various systems, of various sophistication, from Excel spreadsheets to OpenSalt. In the time of Big Data, the expectation is not only that one has access to large volumes of data, but also that the data can be aligned and analyzed on different dimensions in real time – including various item features like content standards.

I am proposing that we rewrite the taxonomies and standards as mathematical vectors, and that we add these vectors as dimensions to the “data cube.” Similarly, we should vectorize the items’ metadata and/or item models and align them on different dimensions of the “cube.” The idea of a “data cube” evolved over time, but the paradigm is easy to communicate and describe. Psychometricians and data scientists can interactively navigate their data and visualize the results through *slicing, dicing, drilling, rolling*, and *pivoting*.

Obviously, the “data cube” is not a cube, given that the different data-vectors are of different length. A data cube is designed to organize the data by grouping it into different dimensions, indexing the data, and precomputing queries frequently. Because all the data are indexed and precomputed, a data cube query often runs significantly faster than the standard queries. Once a data cube is built and precomputed, intuitive data projections can be applied to it through a

number of operations. Also, the traditional psychometric models can be applied at scale and in real time in ways in which was not possible before.

At ACT we are building a Learning Analytics Platform (LEAP) for which I am proposing an updated version of this data-structure: the in-memory database technology that allows for newer interactive visualization tools to query a higher number of data dimensions interactively. In this presentation I will use large-scale examples to illustrate possible alignments based on machine learning tools across multiple testing instruments taken by millions of students.

Psychometrics and response times: What, why, how, and where?

Dylan Molenaar

Thursday, September 13, 2018; 14:45 - 15:45, Room: FORUM

The idea of response times as an important variable in psychological measurement dates back to Francis Galton (1869, 1883) who tried to assess intelligence using the time that subjects needed to respond to a basic stimulus. Due to the lack of methods to accurately assess and statistically analyze the response time data at that time, Galton's idea did not receive a lot of attention (see e.g., Jensen, 2002). Nowadays, methods to record and analyze response times are generally well available. However, interestingly, although response times have been one of the key focusses in mathematical psychology for many decades already (see e.g., Luce, 1986 for an overview), the interest by psychometricians has been relatively limited. Only recently, interest has grown in adding the item response times to the traditional psychometric analysis of the item responses. Currently, many new methods are being proposed that extend existing psychometric tools (like the two and three parameter item response theory models) to include the item response times as an additional source of intra- and inter-individual differences. With these response time methods in place, question arises what the actual contribution is of the response times to the measurement of psychological and educational constructs.

In the present talk this main question is addressed by discussing the historical background, the objective, the methods, and the future of response time modeling on the basis of the What, Why, How, and Where of psychometrics and response times:

- What brought psychometrics and response times together?
- Why are psychometricians interested in adding the response times to the analysis of the responses?
- How can we extract the desired information from the response times using the current state of the art psychometric approaches?
- Where does the scientific study of response times ultimately bring us?

Panel Discussion

Educational measurement and educational theory: two fields apart?

Chair: Prof. Sigrid Blömeke, CEMO director, University of Oslo, Norway

Wednesday, September 12, 2018; 15:45–17:15, Room: FORUM

This panel discussion intends to evoke a lively conversation on the connection between educational measurement and educational theory. One could argue that these are two fields widely apart in everyday research but also look at them as two sides of the same coin. In addition, the answer to this question may depend on a researcher's background. The panel includes therefore a wide span of expertise, representing educational theory, educational policy and learning sciences, applied qualitative and quantitative research as well as educational measurement, assessment, psychometrics and statistics.

Examples of questions to be discussed are

- Can education measurement exist without educational theory?
- Can education theory exist without educational measurement?
- What role does educational measurement play in the evolution of educational theory?
- What role does educational theory play in the evolution of educational measurement?
- Are both theory and measurement required to develop an understanding of educational phenomena?

The panel debate takes place in the Oslo Science Park (Forskningsparken), Gaustadalleen 21, Oslo, Norway, and is open for everybody.

Panelists

- **Alina A. von Davier**, Senior Vice President at ACT and Adjunct Professor at Fordham University, USA. The focus of Davier's research is on the development and application of computational psychometrics, in particular on blending machine learning algorithms with psychometric theory.
- **Cees Glas**, Professor at the Department of Research Methodology, Measurement and Data Analysis of the Faculty of Behavioural Science at the University of Twente, The Netherlands. The focus of Glas' research is on estimation and testing of latent variable models, in particular of IRT models, as well as on the application of IRT models in educational measurement.
- **Gabriele Kaiser**, Professor for mathematics education at the Faculty of Education of the University of Hamburg. The focus of Kaiser's research is on teacher education and teachers' professionalism, modelling and applications in school, international comparative studies, as well as on gender and cultural aspects in mathematics education.
- **Eckhard Klieme**, Director of the Department of Educational Quality and Evaluation at the German Institute for International Educational Research in Frankfurt/M., Germany. The

focus of Klieme's research is on school effectiveness and teaching quality, assessment of student achievement and on international comparisons.

- **Sten Ludvigsen**, Professor in learning and technology at the Department of education at the Faculty of Educational Sciences of the University of Oslo, Norway. The focus of Ludvigsen's research is on learning and technology in education and work, in particular on learning analytics and theory of science.
- **Monica Melby-Lervåg**, Professor at the Department of Special Needs at the University of Oslo, Norway. The focus of Melby-Lervåg's research is on language and reading development, second language learners, cognitive development and development of math skills. She has conducted several large scale longitudinal studies and randomised controlled trials and is also particularly interested in meta-analysis.
- **David Rutkowski**, Professor with a joint appointment in Educational Policy and Educational Inquiry at Indiana University, USA. The focus of Rutkowski's research is on educational policy and technical topics within international large-scale assessment and program evaluation, in particular how large scale assessments are used within policy debates.
- **Sigrid Blömeke,** CEMO director and Professor of Educational Assessment at the University of Oslo, Norway. The focus of Blömeke's research is on the relationship of competencies and performance, primarily with respect to higher education graduates but also across the life span.

Paper Sessions

An ATA Model for Multistage Testing

Angela Verschoor

Session 1A, 10:30 - 12:00, HAGEN 2

Despite an already long tradition in Multistage Testing (MST), the construction of one still remains an art: decisions regarding stages, composition of modules and routing that have to be taken are usually based on simple rules of thumb, gut feelings or previous experience. On the other hand, Automated Test Assembly (ATA) provides an excellent framework for many decisions to be optimized in a systematic way: which combination of items fulfills all specifications but still provides the most accurate measurement? Unfortunately, all ATA models devised until now only regard linear tests.

Even for relatively simple situations, questions like “At what length of the first stage in a two-stage test will the measurement error be optimal?” will yield varying answers from experts, while clearly only one answer could be correct.

In this paper, we present an ATA model for MST. The model user only needs to specify a limited set of specifications: next to the “standard” requirements for linear testing (content restrictions, practical considerations, etc.), the model assumes only an outline of the desired MST design: a number of stages, and a number of modules per stage). The other decisions (selection of items into the modules, routing rules) will be optimized in the model. For the objective function, two possibilities are offered: the first objective function assumes a flat threshold for the Fisher information function over a user-defined interval, while the second objection function minimizes the Root Mean Squared Error for a target population. As the model is non-linear, standard LP-approaches to solve these models might be cumbersome. Therefore, local search methods like Genetic Algorithms or Simulated Annealing seem to be more appropriate for this class of models. A very simple local search method will be presented, providing optimal or near-optimal results in short time.

Although results are heavily dependent on the exact constraints and available item pool, the model shows that in general in a two-stage test a relative short first stage will outperform a test with a longer first stage. Similarly, a 1-3-3 MST will in general outperform a 1-2-4 MST.

Peculiar Subgroup’s Aberrance Response Behavior in Multistage Adaptive Testing: A simulation study

Yuan-Ling Liaw

Session 1A, 10:30 - 12:00, HAGEN 2

The purpose of this simulation study is to investigate peculiar subgroup's aberrance response behavior under a 3-stage multistage test (MST) design. Aberrant responses may lead to proficiency estimation error because the estimates would not reflect the examinees' actual proficiency. Topics related to the examinees' aberrant responses, such as person-misfit statistics, item selection strategy, and response time, have been widely investigated under the computerized adaptive test (CAT) context (Karabatsos, 2003; Meijer, 2003; van der Linden, 2008). Like CAT, MST is also adaptive. However, MST differs substantially from CAT in terms of its design structures and adaptive algorithm. MST utilizes routing decisions that are based on performance on a series of preassembled test items, called *modules*. A test form consists of a series of *stages* in which one or more modules are administered. An MST design consists of a small number of separate modules, and each module can be assembled to meet a set of specifications such as item content and item difficulty. Adaption to an examinee's ability occurs between stages of the testing process and is based on the examinee's cumulative performance on previous item sets. Accordingly, fewer adaptation points are available under MST. MST designs vary substantially as a function of numbers of stages, numbers of modules, or numbers of items in each module. Figure 1 shows an example of a three-stage multistage testing structure. Hence it is difficult to generalize the findings derived from CAT directly to the MST context. Few studies have investigated aberrances in examinee behavior in MST and only two-stage design was investigated (Kim & Moses, 2016). As MST has received attention for their features and efficiency nowadays, more research on examinees' aberrant responses in MST literature is needed.

The simulation is based on a three-stage 1-2-3 MST with each item parameterized according to the two-parameter logistic (2PL) item response theory (IRT) model. For the no aberrance condition, the average of item difficulty parameters was set to be 0.00 for Stage 1, -0.5 for low and +0.5 for high at Stage 2, and -1.0 for low, 0 for middle, and +1.0 for high at Stage 3. The averaged item discrimination parameters are set at either 1 or 0.5 in all modules. I further simulate 100 examinees at each of 41 quadrature points on a theta scale ranging from −3.0 to +3.0, with an interval of 0.15 (N =4,100). To simulate peculiar subgroup's item responses which differ from the no aberrance condition, item responses are generated using fake item parameters to manipulate the levels of difficulty and discrimination, i.e., 15% of the examinees at each theta point; (1) subtract 0.3 from the true *b* parameters; or (2) increase 0.3 from the true *b* parameters; or (3) subtract 0.1 from the true *a* parameters; or (4) increase 0.1 from the true *a* parameters. The achievement estimates are compared with their true proficiency means (i.e., generated thetas). Full details are omitted here due to space constraints. The present findings are expected to contribute to the MST literature.

Routing in the Multistage End of Primary School Test

Maaike M. van Groen, J. Hendrik Straat, & Marie-Anne Keizer-Mittelhaëuser

Session 1A, 10:30 - 12:00, HAGEN 2

Major changes are currently made on the Dutch end-of-primary-school placement test. The main change consists of making new multistage tests available. Per 2018, a domain-specific multistage

test is constructed following a 1-3-3 test design. All students will make the same first stage module. After this initial module, a routing decision is made. One of three modules is selected based on the student's previous responses. After finishing the second module, a similar routing decision will be made. This implies that routing will take place twice for each of the three subjects. Given that six routing decisions will be made per student and the influence of routing on test results, it is important to consider routing carefully.

Many routing methods have been described in the literature. Several of these methods were investigated for this specific test using simulations. One challenging aspect about routing is that some methods also need prespecified cutoff points. For example, when raw scores are used as routing cutoff points one needs to determine their precise value. The student's raw score is then compared with the cutoff points to determine the next module for the student. Although the routing method itself is simple and easy to compute, the method for specifying the cutoffs is more complicated. This specification can be done in a number of ways. One option is to specify the cutoffs such that equal proportions of students are routed through each path. Another option is to use simulations to determine the optimal cutoff points. Optimality is then determined using a criterion such as the precision of the reported ability estimates, the proportion of students per test path, or the classification accuracy. This implies that depending on the routing method, a second method can be required for specifying the cutoff points.

Depending on the routing method choices can be made regarding the input for the routing method. Are decisions based on all previously administered modules or on the last administered modules? Is maximum information computed for the next module or all remaining paths? Are all possible paths admissible? We will discuss the considerations for different choices and use simulations to demonstrate the effect of different choices.

Before the first test administration simulations were run to investigate the many choices that could be made for routing. Decision making about routing was supported by those simulations. After the first test administration in April 2018, the effects of those decisions will be evaluated. How many students took each path through the test? Which changes should be made in the routing procedure for the 2019 test administration? How can we further improve routing? These and other questions will be reflected upon based on the data from the 2018 test administration.

A Feasibility Study of Multistage adaptive design in classroom assessments

Yan Bibby

Session 1A, 10:30 - 12:00, HAGEN 2

In computer adaptive testing, the multistage design with branching rules at decision points has seen some important applications in recent years. A feasibility study of a multistage computer-based adaptive test was carried out in 2016 as part of a system-wide program to be used in schools. The aim of this feasibility study was to evaluate the design and the branching rules. This paper will present the design, the implementation and the evaluation results.

The adaptive test was designed for three grade levels covering four different domains. The adaptive design is a three-stage tailored test design with a specified branching model with two decision points. Each test module consisted of a total of six mutually exclusive testlets with varying average testlet difficulties. Each participating student was administered three testlets, one at each stage, assigned according to the branching rule based on the student performance up to the decision point. All items used in the multistage adaptive test were calibrated based on data collected previously from traditional paper tests using the Rasch model using ACER ConQuest. Test items were assigned to the testlets and branching rules were developed based on these existing item difficulties.

The test data in the feasibility study were analysed, test items were recalibrated. Item difficulties and student abilities were estimated on a common scale. The cut scores predefined in the branching rules were evaluated. The result shows that most of the cut scores worked well, a few of them needed a small adjustment (within ±2 raw score points).

The recalibration results show that the difficulty of testlet in stage 1 which was administrated to all students is approximately the average of all testlets together. The test characteristic curves show each of five paths are well separated from each other for the most part of the ability range. The student proficiencies were estimated using weighted maximum likelihood (WLEs). Student ability distributions were compared by test path. The multistage adaptive test had functioned well in terms of assigning students of different ability levels to the appropriate test paths. The data had clearly showed the branching rules had successfully directed students of higher abilities to the more difficult test paths and students of lower abilities to the easier test paths. There were marked difference in the range of student abilities between students who were assigned different paths at the end of stage 1, and the tailored test design further separated the students into different test paths based on ability at the end of stage 2.

Can Multistage Testing Bridge the Cultural Measurement Divide?

Leslie Rutkowski

Session 1B, 10:30 - 12:00, VIA

Due to language, geographical, or other cultural differences, international measurement across many dozens of participants is a marked challenge, theoretically and operationally. A foundational problem is to ensure that measured constructs are equivalent, particularly as tests are increasingly tailored for groups of countries (e.g., easy booklet design in PISA). Although empirical evidence from recent rounds of one international survey – PISA – have shown a high degree of measurement equivalence across participating countries and educational systems (OECD, 2017), concerns persist over whether a common scale can be used to measure everyone (Kreiner & Christensen, 2014; Rutkowski, Rutkowski, & Liaw, 2017). And in less economically developed countries, technology is a barrier, as PISA and other assessments move to a computerized platform. This is all the more prescient as the OECD weighs moving toward a multistage adaptive

test (ETS, 2016), as this innovation offers promise and peril. A key advantage of a multistage adaptive test (MAT) is the possibility for the test to be more precisely targeted toward the test takers' proficiency, while also limiting the operational burden that is typically associated with fully adaptive tests. However, testing organizations must balance this benefit against potential risks to trend measurement, cross-country comparability, and stable parameter estimates. In the current paper, I address these issues in the context of meaningful and expanding cross-cultural measurement variation. In particular, I discuss what can reasonably be gained by a MAT when countries vary widely in proficiency; the degree to which existing (trend) item banks can be brought to service; and whether items with characteristics typical of past PISA cycles can fulfill future MAT needs. I take both an empirical and simulation-based perspective to highlight several critical issues, should a MAT be adopted for upcoming rounds of PISA and in new PISA instantiations (e.g., PISA for Development).

Problem solving and its role in large-scale assessments: Transversal skills in educational research

Samuel Greiff

Session 1B, 10:30 - 12:00, VIA

One goal of society is placing people in jobs and educational tracks according to their individual skill level and systematically fostering their abilities. To do so, these skills have to be quantified in one way or the other. This talk considers problem solving and several types of it (e.g., Adaptive Problem Solving, Complex Problem Solving, and Collaborative Problem Solving). In fact, problem solving plays an important role in recent large-scale assessments as transversal skill including PISA and PIAAC. For instance, a computer-based assessment of Complex Problem Solving was included in the PISA 2012 survey and Collaborative Problem Solving was assessed in the latest PISA 2015 cycle with over half a million students in over 70 countries. While results of these assessments will yield important implications for educationalists and politicians around the globe, the role of problem solving is controversial among cognitive scientists. In this talk, conceptual backgrounds, assessment instruments, empirical findings, and political implications will be presented in a nutshell and directions for future scientific endeavors will be discussed.

Treatment of Missing Covariate Data in the Scaling Model in Large-Scale Assessments

Simon Grund, Oliver Lüdtke, & Alexander Robitzsch

Session 1B, 10:30 - 12:00, VIA

In educational large-scale assessments (LSA), the method of plausible values (PVs) is used to correct measurement error in the achievement test and to represent students' (latent) proficiency scores while taking covariates from the background questionnaire, such as learning attitudes or interests, into account (Mislevy, 1991). This method follows the multiple imputation (MI) approach of Rubin (1987) by considering the latent proficiency scores as missing data, thus generating

predictions for students' proficiency from a scaling model that is based on both the achievement test data and the covariates in the background questionnaire. However, the scaling procedures employed in the generation of PVs require that the covariates are completely observed. This raises the question of how PVs should be generated from the scaling model when the covariates in the background model contain missing data (Rutkowski, 2011; von Davier, 2013).

In the present talk, we consider different strategies for dealing with missing data in the covariates of the scaling model. This includes the procedures currently employed in educational LSAs such as PISA, which rely on recoding the covariates with missing data before they are entered into the scaling model. In addition, we consider different strategies for treating the missing data that rely on nested and non-nested MI. In this context, non-nested MI refers to procedures that attempt to treat measurement error and missing data simultaneously (i.e., in a single stage), whereas nested MI refers to strategies that generate imputations for missing data and PVs in two consecutive stages (Harel, 2007; Rubin, 2003).

Finally, we present the results from a simulation study that compared these methods in a number of different settings. We show that the procedures currently employed in PISA can lead to biased parameter estimates when the data are not missing completely at random. By contrast, nested and non-nested MI are shown to provide unbiased estimates even with systematically missing data. In addition, we show that simplified procedures on the basis of nested MI which use only a single imputation in the second stage can provide similar results without the need for specialized software implementing the pooling methods required for nested MI. In this context, we emphasize the important differences in perspective of those involved in the scaling of the achievement data on the one hand and those performing secondary analyses on the basis of PVs on the other hand. We close with a discussion of our findings and consider possible consequences for current and future practices of handling missing covariate data in the scaling model in educational LSAs.

De-weighting Plausible Values in International Large-Scale Assessment: A new method for reducing measurement variance.

Eva de Schipper, R.C.W. Feskens, G. Maris, & I. Partchev

Session 1B, 10:30 - 12:00, VIA

International large-scale (educational) assessment (ILSA) studies, such as PISA, measure the abilities of students across countries in different subjects such as mathematics or reading. The main goal of ILSA studies is to make accurate inferences about populations. For various reasons, it is common in ILSA studies to administer different sets of questions to the sampled students. Expressing this additional degree of uncertainty of measurement for an individual is one of the reasons that ILSA studies usually provide several possible scores for each individual, called plausible values. These are used to estimate the statistics of interest as well as to estimate the amount of measurement error in said statistic. Among the usual statistics of interest, especially proportions suffer from a larger amount of measurement error due to the granular nature of plausible values; they are either above or below a specified standard, essentially dichotomizing the

continuous plausible value. The measurement error is further inflated once the plausible value for an individual is multiplied by their sampling weight.

In this study, a method called *de-weighting* is introduced to prevent this inflation of measurement error in ILSA population estimates due to sampling weights. In short, the de-weighting of plausible values entails imputing as many plausible values for a sampled individual as their sampling weight, as opposed to simply multiplying a single imputed plausible value by their sampling weight. The variability of the individuals in the population who are not sampled and therefore represented through a sampling weight is thus taken into account. The findings suggest that the proposed method greatly reduces measurement variance, thereby also decreasing the standard error of the estimates. As hypothesized, the effect is greatest when estimating proportions. The possible implications of using the de-weighting method for ILSA studies are discussed.

A quadrature Kalman filter for estimating MIRT models for sequential data

Peter van Rijn

Session 2A, 12:45 - 14:15, HAGEN 2

Although numerous approaches exist for modeling sequential data in the context of item response theory (IRT; von Davier, Xu, & Carstensen, 2011), some issues still persist. A first issue concerns the sequential nature of the data. Some approaches that have been applied in a longitudinal context are problematic because they do not take into account the order of the observations. For example, the model described by te Marvelde, Glas, Van Landeghem, and Van Damme (2006) and Andrade and Tavares (2005) assumes a straightforward multivariate normal distribution for the latent variables at different measurement occasions. However, in this approach, any permutation of the measurement occasions will show the exact same fit, because the order of the observations is not taken into account.

A second issue that can be distinguished is computational tractability. Some approaches can relatively quickly become computationally intractable, because at each time point a new latent variable is introduced (Embretson, 1991; Fischer, 1989; Andrade &Tavares, 2005; te Marvelde et al., 2006). This increases the dimensionality proportional with the number of time points, and is generally referred to as the curse of dimensionality. The problem lies in the estimation, which in IRT typically is performed by maximizing the marginal likelihood. This likelihood is obtained by integrating out the latent variables. Although current estimation methods in IRT can deal with higher dimensions, in practice, the effective dimensionality of the integral (after dimension reduction techniques) cannot be larger than six or so (Cai, 2010). This problem becomes even more pertinent when a multidimensional IRT model is to be used at each time point (Rijmen, 2010; Cho, Athay, & Preacher, 2013).

A general method for estimating multidimensional item response theory (MIRT) models for longitudinal and time series data, which addresses both these issues, is presented. The method employs an expectation-maximization (EM) algorithm in which the expectation step is formed by a

discrete-time Kalman filter that makes use of adaptive Gauss-Hermite quadrature to deal with nonlinearity and non-normality (Arasaratnam, Haykin, & Elliott, 2007). The use of quadrature is highly similar to marginal maximum likelihood estimation of regular MIRT models, thereby providing a natural extension of the latter method to longitudinal and time series settings. Two applications of the method to real educational data are discussed.

Efficient estimation of item response theory models with multiple groups in large-scale educational assessments

Björn Andersson

Session 2A, 12:45 - 14:15, HAGEN 2

In large-scale educational assessment programs, students in different countries or regions are assessed in subject domains such as mathematics, reading and science. In these programs, the underlying model is an item response theory model which defines the relationship between a hypothesized latent variable vector, the background variables and the observed item responses. Due to the large amount of data involved in these assessment programs, several simplifying assumptions are usually made when estimating the parameters of the underlying model. These assumptions include a unidimensional latent variable and measurement invariance across regions. This talk presents a new estimation method using a second-order Laplace approximation of the likelihood for multidimensional multiple group item response theory models which enables the use of more realistic models in large-scale educational assessment programs. We illustrate how the proposed method can be used to improve the estimation of population parameters using large-scale assessment data and suggest ways in which the operational procedures can be modified to better assess the performance of students in individual countries or regions.

Nonlinear item-level moderation in measurement models: Exploring the relationship between product and process data

Maria Bolsinova & Dylan Molenaar

Session 2A, 12:45 - 14:15, HAGEN 2

When educational tests are presented in a computerised form, it is feasible to not only record the product of the response process (i.e., response accuracy or response choice), but also the characteristics of the process itself. For each combination of the person and the item different values of many additional variables could be recorded: response times, confidence ratings, verbally reported response processes, number of actions in interactive items, number of item clicks, number of eye fixations on the areas of interest, inspection times, response changes, certainty scores, or physiological measures. These variables can be included as moderators in the measurement models for the ability of interest such that one can investigate whether the probability of a correct response is related to the value of the moderator and whether there is an interaction effect between the measured ability and the moderator. For moderators that vary across persons but not across items (e.g., traditional moderators like age or SES) there is a wide

variety of multi-group, linear, nonlinear and nonparametric methods for investigating these effects. Item-level moderators have received much less attention in the latent variable model literature. The development of methods to test for interactions between item-level moderators and the ability has recently started to evolve across similar lines as in traditional moderation models. That is, approaches have been proposed that require categorisation (Partchev & De Boeck, 2012) of the item-level moderators and models have been proposed by specifying linear functions between the intercept and slope parameter of the measurement model and the item-level moderator (Bolsinova, Tijmstra, & Molenaar, 2017; Goldhammer, Steinwascher, Kroehne,& Naumann, 2017).

However, parametric nonlinear and nonparametric models for indicator-level moderation are lacking while such approaches are valuable in exploring the exact form of the relationship between the moderator and the parameters of the model. The assumption of linearity of item-level moderation might be violated in practice, and using linear models might lead to invalid conclusions about the relationship between the parameters of the measurement model and the item-specific moderator. For instance, one might conclude that the intercept increases with the values of the moderator (e.g., that slower responses on a science test are more often correct), while it might be that it increases only up to some value of the moderator and decreases after that value, or that the increase is not linear. Therefore, we propose to model the relationship between the item-specific moderator and the parameters of the measurement model in a more flexible way. In this presentation, parametric nonlinear and nonparametric item-level moderation methods are developed. In a simulation study we demonstrate the viability of these methods. In addition, the methods are applied to a real dataset pertaining to arithmetic ability in which the main and interaction effects of response time are investigated.

Revisiting the Bahadur representation of sample quantiles for the standard error of kernel equating

Gabriel Wallin & Jorge Gonzalez

Session 2B, 12:45 - 14:15, VIA

In educational measurement, the comparability of test scores coming from different test versions is essential for fair assessments of test takers. Statistical models for test score equating enable such comparisons using a functional parameter–the equating transformation–that maps the scores from the scale of one test form into their equivalents on the scale of another. The standard error of equating (SEE) is one of the most common measures used when evaluating an equating transformation. In kernel equating (KE), the SEE has been derived using the delta method (von Davier, Holland, & Thayer, 2004), by relying on the asymptotic normality of the maximum likelihood estimators of score probabilities. Thus, using the delta method for calculating the SEE would be theoretically valid only when estimated score probabilities are, at least, approximately normally distributed. Because score probabilities in KE are commonly estimated after presmoothing the score distributions using maximum likelihood estimates from loglinear models, this issue has not been of great concern.

In recent years, however, alternative methods of presmoothing have been suggested. Some of them do not necessarily lead to approximately normally distributed estimated score probabilities, which makes it possible to question the validity of the current method for calculating the SEE. An alternative method that does not rely on normality and that use the Bahadur representation of sample quantiles (Bahadur, 1966) was suggested by Liou, Cheng, and Johnson (1997). These authors derived SEE expressions for the nonequivalent groups with anchor test designs, considering both equating estimators that used the Gaussian kernel as well as the uniform kernel for the continuization of the score distributions. To the best of the author's knowledge, no comparison has been made between the delta method of computing the SEE and the method that uses the Bahadur representation of sample quantiles. In this study, the alternative method suggested by Liou, Cheng, and Johnson is expanded to: i) include expressions for the SEE that can be used regardless of the choice of the kernel function, ii) include other data collection designs (e.g., the equivalent groups design), and iii) obtain an expression of the SEE for the chained equating transformation. With these new results at hand, the two different methods of calculating the SEE are compared for different data collection designs, kernel functions, presmoothing models, and using both post-stratification equating and chained equating.

A new likelihood approach for the simultaneous estimation of IRT equating coefficients on multiple forms

Waldir Leoncio & Michela Battauz

Session 2B, 12:45 - 14:15, VIA

Test equating is a statistical procedure to ensure that scores from different test forms are comparable and can be used interchangeably (González and Wiberg, 2017). Within the Item Response Theory framework, if the statistical modeling of each test form is performed independently, their respective parameters will be on different scales and thus incomparable. Equating solves this problem by transforming item parameters so they are all on the same scale. Popular methods for equating pairs of test forms include the mean-sigma, mean-mean, Stocking–Lord and Haebara (Kolen and Brennan, 2014). For multiple forms, it might be necessary to employ more elaborate methods which take into account all the relationships between the forms.

We are proposing a new statistical methodology that simultaneously equates a large number of test forms. Simultaneous equating methods are not new in the literature, with Haberman (2009) proposing a linear regression method, Battauz (2013) presenting chain and average equating coefficients and Battauz (2017) introducing the generalization of some well-known methods such as those mentioned above. Our proposal differentiates itself from the current state-of-the-art by using the likelihood function of the true item parameters and the equating coefficients to perform the concurrent estimation of all equating coefficients. By taking into account the heteroskedasticity of the item parameter estimates as well as the correlations between the item

parameter estimates of each test form, this new method yields equating coefficient estimates which are more efficient than what is currently available in the literature.

When dealing with large-scale assessments, often composed of several test forms with dozens or hundreds of items each, the number of parameters to be estimated can easily become a concern. After all, each new item adds at least one IRT parameter to the likelihood function, and any additional test form can introduce several new items as well as two mandatory equating coefficients. This can quickly make the proposed approach too complex from a computational point of view. We overcome this problem by considering the equating coefficients as parameters of interest and the true item parameters as nuisance parameters. With this setup, the profile likelihood can be used instead of its complete counterpart, thus potentially saving the costly estimation of hundreds of parameters of secondary importance.

The statistical and computational properties of the methods developed are being investigated under controlled simulations and the results are promising. Possible practical applications include any large- scale assessment which administers and equates several test forms.

Criterion-referenced adaptive university exams: Effects of different linking designs on ability estimates

Aron Fink, Sebastian Born, Andreas Frey, & Christian Spoden

Session 2B, 12:45 - 14:15, VIA

The increasing digitalization in the educational sector opens up new opportunities not only for the teaching process, but also for the design of written university exams. Digital technologies make it possible to use innovative item formats and have the potential to foster the efficiency of scoring and data handling. Furthermore, and even more important from a scientific point of view, the shift to using digital technology for testing purposes in higher education provides the opportunity to implement state-of the art-methods from Psychometrics and Educational Measurement in the day-to-day practice. In particular, criterion-referenced computerized adaptive testing (CR-CAT) has the potential to make university exams more individualized, more accurate and fairer. From a practical point of view, however, the calibration of the item pool needed for CR-CAT poses a critical challenge since a separate calibration study is often not feasible and/or sample sizes of university exams are too low to allow for a stable estimation of item parameters. Thus, we suggest a new method for continuous item pool calibration during the operational CR-CAT phase. This method enables a step-by-step build-up of the item pool across several time points without a separate calibration study. In order to keep the scale constant across time points, link items are used. Due to the novelty of the method, the impact of the proportion of link items used and their item difficulty distribution on the quality of the person ability estimates (q) is unclear. To shed light into this, a simulation study based on a fully crossed design with the four factors “proportion of link items” (1/6, 1/4, 1/3 of test length), “difficulty distribution of link items” (normal, uniform, bi-modal with very low and very high difficulty only), “test length” (36, 48, 60 items), and “sample size” per time point (50, 100, 300) was carried out. Evaluation criteria for the quality of the q

estimates are the bias conditional on q and the standard error of q conditional on q. The study is currently running, but will be completed before the conference. Regarding the results, we expect that a higher proportion of extremely difficult link items will reduce both bias and standard error for persons at the margins of the ability distribution. Longer test lengths and larger sample sizes should lead to less bias and lower standard error for all persons.

Resetting the standard with IRT concurrent calibration and Circle-arc equating in small samples

Monika Vaheoja, Norman. D., Verhelst, & Theo. J. H. M. Eggen

Session 2B, 12:45 - 14:15, VIA

Resetting standard performance statistically in tests with small samples is challenging because the small sample statistics often include bias, caused by sampling error. In practice, therefore, are the standard setting procedures applied that rely on experts' estimation such as Angoff (1971), and empirical information to statistically reset the standard is neglected. But the standard-setting methods that include experts estimation, are biased too, and often expensive (Cizek & Bunch, 2007).

Livingston and Kim (2009) proposed a circle-arc equating for small samples. This method assumes a curvilinear relationship between reference and a new test to prevent the transformation of the scores beyond the range of possible scores. Different studies have shown promising results in favor of circle-arc method (Dwyer, 2016; LaFlair, Isbell, May, Gutierrez & Jamieson, 2015), but because the circle-arc method is a solution from the classical test theory approach it has its limitations too. Especially in the context where the population ability and test difficulty interact. In the later, Item Response Theory (IRT) outperforms classical test theory, but until now, it is not advised for small samples (Kolen & Brennan, 2014).

IRT is a theory about the responses of participants on a given test or exam. In this theory, the probability of correctly answering an item by a respondent is modeled assuming that a score on an item is dependent on the ability of the respondent and of the item characteristics. One of the IRT models is the One Parameter Logistic Model (OPLM; Verhelst & Glas, 1995). In OPLM are the item difficulty parameters estimated with the conditional maximum likelihood estimation, which means that no assumption of the population ability has to be made and the sample does not have to be representative to the population.

In the present simulation study, we will compare Circle-arc equating and IRT concurrent calibration with OPLM in transferring cut-score from reference test to a new test in three different contexts: at first we will fix the reference sample during the calibration, second, set them free and in the third context we will vary the population ability on a new test with low, and high ability group. Within each context, data is simulated in three varying situations: sample size, test length, and test difficulty. The results demonstrate that even in small samples (50 subjects taking

both tests) IRT method outperforms classical test theory approach when tests' difficulty and population ability interact. The discussion involves the suggestion for further research such as the influence of the anchor-test and the reliability of the tests in the equating.

Measurement invariance in PISA 2015: A systematic investigation of patterns across questionnaires, scales and countries

Janine Buchholz & Johan Braeken

Session 3A, 9:45 - 11:15, HAGEN 2

International large-scale assessments (ILSAs) such as the *Programme for International Student Assessment* (PISA), *Trends in International Mathematics and Science Study* (TIMSS) and *Progress in International Reading Literacy Study* (PIRLS) aim at measuring and comparing latent constructs between respondents from a large number of participating countries -- an endeavor which requires measurement invariance (MI) across all participating countries to be established. The most commonly employed technique for MI testing is multigroup-CFA (MGCFA; e.g. Greiff & Scherer, 2018). Yet, the method was proven unsuitable given the large number of countries participating in these assessments (Rutkowski & Svetina, 2014). In addition, it capitalizes on global model fit, thus being unable to point at group-specific misfit.

Using a recently developed measure of group fit rooted in MGCFA, the present study presents a systematic investigation of the 58 questionnaire scales reported in the most recent cycle of PISA (OECD, 2016) for the following reasons: (1) PISA can be regarded as having "strategic prominence in international education policy debates" (Hopfenbeck et al., 2017, p. 1); (2) with about 70 participating countries in PISA 2015, the number of tested groups is particularly large; (3) within ILSAs, the questionnaires are hardly ever subject to MI testing (e.g. Braeken & Blömeke, 2016); (4) most scientific publications on PISA focus on secondary analyses of constructs administered with the questionnaires (Hopfenbeck et al., 2017), thus placing an operational need on the appropriateness of comparisons across countries in these studies; (5) in its most recent cycle, PISA implemented an innovative approach for MI testing using IRT item fit (OECD, 2016), thus raising the question about the replicability of their findings in the context of more common analysis techniques.

Based on a quantification of the amount of measurement (non-) invariance across scales and countries, we will report on identified patterns due to scale properties (e.g., length, response categories, previous use) and country characteristics (e.g., previous participation, geographic location, language groups, gross domestic product). These findings will help to identify country subsets for which meaningful comparisons are appropriate, and they may also be used to guide questionnaire development in the context of ILSAs.

Detecting differential item functioning with entropy in logistic regression

Brandi A. Weiss & William R. Dardick

Session 3A, 9:45 - 11:15, HAGEN 2

In this talk we will discuss the adaptation of four entropy variants to detect differential item functioning (DIF) in logistic regression (LR): entropy (*E*), entropy misfit (*EM*), the entropy fit ratio (*EFR*), and a rescaled entropy fit ratio (Rescaled-*EFR*). Logistic regression is frequently used to detect DIF due to its flexibility for use with uniform and nonuniform DIF, binary and polytomous LR, and groups with 2+ categories. In this talk we will focus on binary LR models with two groups (reference and focal), however, we will also discuss the use of entropy with polytomous LR models and models with 2+ focal groups. We will present both a mathematical framework and results from a Monte Carlo simulation.

A fair test is free of measurement bias and construct-irrelevant variance. When groups are found to differ on an underlying construct test fairness may be impacted. DIF may help identify potentially biased items. While traditionally, dichotomous measures of statistical significance have been used to detect DIF in LR (e.g., c2 and G2), more recent work has emphasized the importance of simultaneously examining measures of effect size. Model fit statistics can be thought of as a type of effect size. Previously, entropy has been used to capture the separation between categories and is expressed as a single measure of approximate data-model fit in latent class analysis, data-model fit in binary logistic regression, person- misfit in item response theory (IRT), and item-fit in in IRT. Entropy captures discrimination between categories and can be thought of as a measure of uncertainty that may be useful in conjunction with other measures of DIF. In this presentation we extend entropy for use as a measure to detect DIF that complements currently utilized DIF measures.

Monte Carlo simulation results will be presented to demonstrate the usefulness of entropy-based measures to detect DIF with a specific focus on model comparison and changes in entropy variants. We evaluate the following variables across 1,000 replications per condition: sample size, group size ratio, between-groups impact (i.e., difference in ability distributions), percentage of DIF items in the test, type of DIF (uniform vs nonuniform), and amount of DIF. Results will be presented comparing entropy variants to current measures used to detect DIF in LR (e.g., c2, G2, DR2, difference in probabilities, and the delta log odds ratio). Statistical power and Type I error rates will be discussed.

Entropy-based measures may be advantageous for detection of DIF by providing a more thorough examination of between-group differences. More specifically, entropy exists on a continuum thus representing the degree to which DIF may be present, does not rely on dichotomous hypothesis testing, has an intuitive interpretation because values are bounded between 0 and 1, and can simultaneously be used as an absolute measure of fit and a relative measure for between-groups comparisons

The MIMIC pure anchor method for DIF: Detecting psychological impact, not bias

Maryam Alqassab & Gavin T. L. Brown
Session 3A, 9:45 - 11:15, HAGEN 2

Differential item functioning (DIF) indicates a construct-irrelevant factor (e.g., age, sex, or ethnicity) systematically impacts responding to items. DIF studies are usually carried out with demographic groups rather than with psychological grouping variables that might not be construct-irrelevant. DIF could be consistent with a construct that is relevant to the phenomenon of interest suggesting impact rather than bias (Zumbo, 1999).

When items are correlated (i.e., factors), DIF may be inflated by the collinearity of items. The pure-anchor technique within multiple-indicator, multiple cause (M-PA) analysis (Shih & Wang, 2009) uses a DIF-free-then-DIF procedure that fixes one item with no DIF as an anchor to reduce the probability of Type I errors in detecting DIF (Wang & Shih, 2010). The iterative MIMIC procedure (M-IT) tests each item within a construct individually and sets as the pure anchor the item which generated the lowest DIF index (Shih & Wang, 2009).

This study uses a multi-dimensional (i.e., 4 factors, 33 items) research inventory (i.e., Student Conceptions of Assessment, version VI; Brown, 2011) and a brief inventory of student interest and self-efficacy in either reading or mathematics. Higher test scores have been associated with the SCoA factor that assessment is for improvement (Brown, Peterson, & Irving, 2009) and when students have greater interest or self-efficacy (‘Otunuku & Brown, 2007). Hence, DIF in favour of students with higher self-efficacy or interest may indicate impact rather than bias.
Participants (N = 799) were Year 9 and 10 high school students in New Zealand. Interest and self-efficacy in reading and mathematics were used as DIF grouping variables. Participants were grouped by interest (high vs. low), self-efficacy (high vs. low), and test subject (mathematics vs reading comprehension), resulting in small reference and the focal groups (n = 180). DIF by interest and self-efficacy was conducted using M-PA for the four SCoA factors in each subject separately. Only one item was used as an anchor and analysis used the WLSMV estimator (Muthén & Muthén, 2010).

Of the 29 items, after fixing the pure anchor, five items in mathematics and eight items in reading had statistically significant Wald test DIF magnitudes. This contrasted positively to the standard MIMIC DIF analysis which found 18/33 items with statistically significant DIF in mathematics and 17 in reading. A Monte Carlo simulation study of 10,000 replications and two groups of 200 using population parameter values (i.e., number of items per factor ranging from 4 to 10, loadings set at either 0.80 or 0.60) akin to the range of regression weights seen in studies with the SCoA, found that except for expected loadings of 0.80 and either 4 or 10 items per factor, the bias in parameter estimation was much greater than 10% (M=47.88, SD=37.78). This indicates that the observed DIF values are highly likely to be over-estimated, even using the M-PA approach. Items with statistically significant DIF aligned with the known effects of self-efficacy and interest on academic achievement, supportive that impact, not bias was present. Further work with the promising M-PA procedure is warranted.

Enhancing the comparability of self-reported knowledge using the overclaiming technique

Hana Voňková, Ondřej Papajoanu, Jiří Štípek, Miroslava Černochová, & Kateřina Králová

Session 3A, 9:45 - 11:15, HAGEN 2

Respondents' self-reports are often employed in educational surveys (e.g. PISA, TIMSS) and are frequently used to compare different groups of respondents (based on country, socioeconomic status etc.). However, serious concerns have been raised about the comparability of such data, which may be hindered by bias – the score differences on the indicator of a construct do not correspond to the differences in the underlying trait or ability. Such differences in reporting behavior are well-documented across cultures or different groups of respondents. One of the potential sources of scale scores distortion is socially desirable responding (SDR), a tendency for some people to self-enhance when describing themselves.

A promising approach to overcome SDR is the overclaiming technique (OCT). The technique asks respondents to rate their familiarity with a set of items from a particular field of knowledge (e.g. astronomy, history, literature). Some of the items (usually about 20%), however, do not actually exist (foils). By using signal detection analysis, the technique allows us to measure respondents' knowledge exaggeration (the overall tendency to report familiarity with both existent and nonexistent items) and accuracy (the ability to discriminate between existent and nonexistent items). Here we investigate the potential of the overclaiming technique to enhance the cross-country comparability of students' self-reported mathematical knowledge. We also investigate the comparability of self-reported ICT knowledge between different groups of students within a single country.

The cross-country analysis has been conducted using the questions on familiarity with mathematical concepts used in PISA 2012 student questionnaire. The data include the observations of 275 904 students in 64 countries and economies. We show that there are significant differences in responding patterns between particular countries, however, we identify similar patterns of responding in geographically and culturally close country-regions. We also validate the overclaiming scores using external variables like PISA math test scores, GDP and public expenditure in education.

Furthermore, we investigated the potential of the overclaiming technique using the questions on familiarity with ICT concepts administered to two different groups of Czech university students (N=374) – one group studying ICT and the other studying educational sciences (non-ICT). The technique has never been used in the area of ICT skills and knowledge before, even though the self-reports of ICT skills are widely used. Surprisingly, ICT students report being almost twice as much more familiar with non-existing ICT concepts than non-ICT students. This could be interpreted that those who believe their knowledge in certain domain to be excellent may be more

prone to exaggerate (self-enhance) their knowledge. The differences in the self-reported familiarity with ICT concepts between ICT and non-ICT students are substantial both before and after the adjustment using the OCT, however, the adjusted results reflect the tendency of ICT students to exaggerate their knowledge and, to a certain degree, decrease the absolute differences between these groups.

More accurate asymptotic standard error formulas for IRT ability estimators

David Magis

Session 3B, 9:45 - 11:15, VIA

Most-known IRT ability estimators under dichotomous scoring (MLE, BME, WLE and robust) have simple and fancy formulas to derive their associated asymptotic standard errors (ASEs). Such ASEs are of primary interest for determining the degree of precision of the ability estimates but also in more specific contexts, such as e.g., CAT stopping rules. However, some of these ASEs were derived under spurious assumptions, or only recently, and are therefore not yet widespread. The purpose of this talk is to present a general and unified approach to derive ASE formulas for a broad class of IRT ability estimators, that encompass the most-known ones. Using mathematical derivations for asymptotic convergence of Taylor series expansion, a general ASE formula is derived and can be immediately applied to any classical IRT estimator. Some surprising results are encountered and discussed. Eventually, the potential usefulness in e.g., CAT context, is outlined.

Advancing Exploratory Cognitive Diagnosis Models for Educational Measurement and Classroom Assessments

Steven Andrew Culpepper

Session 3B, 9:45 - 11:15, VIA

Advances in educational technology provide teachers and schools with a wealth of information about student performance. A critical direction for educational research is to harvest the available longitudinal data to provide teachers with real-time diagnoses about students' skill mastery. Cognitive diagnosis models (CDMs) offer educational researchers, policy-makers, and practitioners with a psychometric framework for designing instructionally relevant assessments and diagnoses about students' skill profiles. Still, methodological challenges prevent the widespread application of CDMs in educational measurement and classroom assessments. This paper considers problems of fundamental problems of identifiability, model selection, and the validation of expert knowledge.

Accurate inferences for CDM model parameters and student classifications require knowledge about the latent processes and attributes students need to succeed on educational tasks. The CDM Q matrix indicates which attributes are needed for each item and is central to implementing CDMs. In most applications of CDMs, content experts specify Q. The general unavailability of Q

for most content areas and datasets poses a barrier to widespread applications of CDMs and recent research accordingly developed fully exploratory methods to estimate Q. However, current methods do not always offer clear interpretations of the uncovered skills and existing exploratory methods do not use expert knowledge to estimate Q. In fact, estimating Q without the use of available expert knowledge may be sub-optimal. Instead, incorporating expert knowledge during Q estimation may enhance interpretation of uncovered attributes and could assist with cognitive theory development. That is, using an exploratory method with expert knowledge may help to identify residual, or unexplained, attributes that are not predicted by cognitive theory. In such cases, exploratory CDM results can be shared with experts and subsequent conversations may serve to refine cognitive theories.

We consider an exploratory CDM framework that directly uses expert knowledge about item features by developing a new model to relate expert knowledge to the Q matrix using a latent, multivariate regression model. We report new sufficient conditions for identifying model parameters that impose fewer restrictions and are more likely to be satisfied in empirical applications. We show how the developed method can be used to validate which of the underlying attributes are predicted by experts and to identify residual attributes that remain unexplained by expert knowledge. We report Monte Carlo evidence about the accuracy of selecting active expert-predictors and present an application using Tatsuoka's fraction-subtraction dataset. Our analyses partially support expert knowledge and we uncovered two additional attributes that were not previously specified by experts. In general, the results of such analyses could be used to validate expert knowledge and shared with experts to determine if the residual attributes describe previously unidentified cognitive skills. We conclude the paper with a discussion of how the exploratory CDM approach can aid educational measurement in practice with particular focus on the settings where the goal is to provide fine-grained assessment of educational interventions a longitudinal setting.

A HYBRID IRT model for test-taking persistence in low-stakes tests

Gabriel Nagy & Alexander Robitzsch

Session 3B, 9:45 - 11:15, VIA

Results of large-scale assessments of student achievement are sensitive to students' persistence in maintaining a constant level of effort and precision over the course of a test. Low persistence is indicated, for example, by item position effects (IPE) that reflect decreases in the probabilities of correct responses being given towards the end of a test. IPEs are commonly modeled on the basis of assessment designs with rotated item positions by means of IRT models in which the items' difficulty parameters are related to their positions in the test. In these models, the strength of this relationship is typically allowed to vary between individuals. Therefore, IRT models for IPEs allow individual differences in IPEs to be related to ability and to covariates. However, a drawback of the commonly used IRT approach is that it assumes that the students' response process does not change across positions.

In this paper we present an alternative representation of test-taking persistence. We assume that students might change their response behavior from an effortful response mode to random guessing behavior. Drawing upon the HYBRID IRT model, we propose a model that can be applied to rotated assessment designs. The suggested model combines a two parameter logistic (2PL) part with a latent class model, whereby the latent classes represent the first item positions in which individuals have changed their response behavior. Latent class membership is expressed as a function of an underlying normally distributed continuous variable that reflects the individuals' switching points to random guessing behavior. This specification enhances the estimation of latent class proportions, and allows for a straightforward assessment of the relationships of switching points with ability and covariates. The model can be estimated with standard software by means of maximum likelihood estimation via the expectation maximization algorithm.

To demonstrate the model's utility, we applied it to a reading comprehension test (with 32 item positions) administered to fifth-grade students (n = 2,774) by means of a rotated matrix design. Compared to the commonly used IRT model for IPEs, the newly proposed model showed a better fit to the data. Results derived on the basis of the proposed model indicated that higher ability was associated with later onset points of random guessing behavior (r = .46). In addition, students' switching points were clearly related to a test of decoding speed (r = .32). The standard IRT model for IPEs did not indicate any relationship between ability and IPEs and revealed a rather weak relationship between IPEs and decoding speed.

These findings suggest that, at least in the area of reading assessments, students' test-taking persistence might be better represented by qualitative changes in response behavior. Under these circumstances, the proposed extension of the HYBRID model provides a promising tool for assessing test-taking persistence and studying its relationships with ability and covariates.

Exploring log files in international large-scale assessments: Methods, practices and tools

Denise Reis Costa

Session 4A, 11:20 - 12:05, HAGEN 2

The new era of large-scale assessments involves the administration of the tests in a computer-based format. Examples in the international surveys which have conducted assessments in this format are: the Programme for International Student Assessment (PISA) and Programme for the International Assessment of Adult Competencies (PIAAC). Beyond collecting the correct/incorrect answers for each item, both assessments also collect the interactions between respondents and the computer testing application during the course of the test administration. Respondents' actions (e.g. starting a unit, clicking in a bottom, time spending until inputting an answer) within the tool are recorded in log files. The Organization for Economic Co-operation and Development (OECD), responsible for both programmes, has released such files publicly since the 2012 cycle of PISA and PIAAC. This work aims to present the potential and limits of such files, the up-to-date literature, as well as examining tools to extract and/or conduct correct

analysis from this data. For this objective, this study will: (a) give researchers an overview of what they can expect for a first analysis as well the existed tools to work with these log files; (b) present a range of methods and practices that have been using such files to address issues relating to test-taking behaviour and strategies followed by respondents when answering to test items.

Educational assessment and model building using process data: Issues of open science and replication

Johannes Naumann, Malte Elson, & Frank Goldhammer

Session 4A, 11:20 - 12:05, HAGEN 2

Data delivered by Large Scale Assessments (LSAs) are not only used to describe student performance, link performance to background variables on the student, school, and system level, and thus inform educational policy. Rather, LSA data is also increasingly being used for theory building in substantive educational and psychological research. One advantage of using LSA data for substantive research is that research grounded on LSAs already addresses many of the problems recently raised concerning the openness and replicability of educational and psychological research ("replication debate"; e.g. Makel & Plucker, 2014), given large samples and cyclic repetition, which can be utilized for a disentangling of exploratory and confirmatory research, or direct replications. As LSAs are increasingly carried out as computer-based assessments (CBAs), this extends to models requiring data on the task solution process, when log files of student behavior can be mined for psychologically meaningful behavioral indicators. Only few attempts however have been made to date to replicate research using LSA process data.

In the present research, PISA 2012 CBA data was used for an attempt to replicate recently published results that had been obtained using data from the (optional) PISA 2009 Digital Reading Assessment (Naumann & Goldhammer, 2017). In these authors' research, a dual-processing account of reading digital text (Shiffrin & Schneider, 1977; Walczyk, 2000) was tested through an examination of items' difficulties and persons' skills effects on time-on-task effects on performance in digital reading, employing a GLMM-framework. Consistent with a dual processing account, the authors found strong positive time-on-task effects in weak digital readers and hard items, while time-on-task effects were negative in easy items and null in skilled digital readers. Thus, negative correlations emerged between random item and person intercepts, and random item and person specific time-on-task slopes, respectively. Also in line with a dual-processing account, items' navigational demands and persons' comprehension skills, modeled as fixed effects, moderated time- on-task effects thus that time-on-task effects were positive especially in weak comprehenders and in tasks with high navigation demands.

These results were only partly replicated in PISA 2012. While the median correlation between person intercepts and slopes across 17 countries participating in both cycles was -.61 in 2009, pointing to much stronger time-on-task effects in weaker digital readers, the corresponding median correlation in 2012 was only -.30. Also, this correlation was lower in 2012 than 2009 in each individual country. The median correlation between item intercepts and slopes was -.61 in

2009, and -.50 in 2012. Similarly, while the median interaction effect between time on task and comprehension skill in 2009 was -0.07, it was only -0.03 in 2012, and the median interaction between time on task and navigation demands was 0.26 in 2012, while it was 0.48 in 2009.

These results indicate that replicability of substantive results that were obtained using LSA- data must not be taken for granted despite large samples and standardized testing procedures. The present results are discussed in the context of changes in the test design that had occurred between the 2009 and 2012 PISA CBAs.

Optimal scores in comparison to sum scores and parametric IRT scores

Marie Wiberg, James Ramsay & Juan Li

Session 4B, 11:20 - 12:05, VIA

Many standardized tests use sum scores, i.e. number of items correct, as a measure of test takers' ability as they are easy to interpret and computationally fast. Sum scores have however, some limitations as they are calculated after a test has been performed and it targets the whole test and not single items. When constructing a test it is instead common to model the items with parametric item response theory (IRT). A well-known problem with parametric IRT is that not all items can be satisfactorily modeled with a parametric IRT model. Recently, optimal scores was proposed to be used in addition to sum scores and serves as a flexible alternative both for scoring the test and for estimating item characteristic curves. In optimal scores, the interaction between test takers' performance and item impact is used, thus giving more weights to items with more information. The aim with this presentation is to present and discuss optimal scores and compare it with sum scores and parametric IRT scores using both real test data and simulated data. Examples of how to fit different real test items will be given in comparison to parametric IRT models. The simulation study will examine bias and root mean squared error for optimal scores as compared with the alternatives. The results indicate that we can improve the accuracy if optimal scores are used and that optimal scores provide a flexible alternative for estimating item characteristic curves. The latter is especially of interest when we have items, which does not fit a parametric IRT models. The presentation ends with a discussion, which include some future direction of research.

On proficiency scales and errors of measurement for educational tests

Svend Kreiner & Jeppe Bundsgaard

Session 4B, 11:20 - 12:05, VIA

Educational tests are used for two main reasons: 1) to give teachers insight into the abilities of their students, and 2) to give administrators, researchers, the public and politicians knowledge of status, progression and relative level of a group of students (as compared to another).

These two reasons put different demands on the tests. In the first case, the teacher want to know with some confidence what an individual student is capable of, knows and understands, and the

teacher welcomes suggestions on how to help the students reach the next goals. In the second case, administrators etc. want to know if progression was made and whether certain thresholds was reached for a specific population.

Combinations of the two goals are possible, but hard to attain. In international large-scale assessments like PISA, and the IEA assessments (PIRLS, TIMMS, ICILS etc.), it is in principle possible to provide teachers with scores on proficiency scales defined by subject matter arguments relating to student progressions, but results on individual students are never reported. At classroom and student levels, test results are often collected at set points in time so that results can be aggregated to provide information to administrators at higher levels, whether or not it is convenient and/or useful for the teacher to have information on the class and the student at the time where administrators need them. Since test results are collected at specific points of time during the school year, it is possible not only to reports simple transformed raw scores, but also percentile scores that can be used to compare test results for separate students to the complete student population, and it is rare to find examples where test results at student levels are more than simple transformed raw scores and percentile scores.

This paper is an argument for development of and use of proficiency scores for applications of educational tests at classroom and student level. We will discuss different ways to interpret test scores and different ways to construct informative proficiency scores (e.g. Fraillon et al., 2015; OECD, 2014; Draney & Wilson 2011; Wilson & Santelices 2017), providing more useful information than transformed raw scores and percentile scores; and we will show how to assess the measurement error of proficiency scores. The methods will be illustrated with data on proficiency scales for a test measuring 21st Century Skills (Bundsgaard, 2018; Bundsgaard, in review), on data from The Danish National Test (DNT), and data from *International Computer and Information Literacy Study* (ICILS 2013) (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014).

Using educational measurement to support countries to achieve the Sustainable Development Goals in post-conflict contexts

Dan Cloney, Alex Daraganov, Leigh Patterson, Ray Adams, Ross Turner, & Maurice Walker

Session 5A, 13:00 - 14:30, HAGEN 2

From 2012 through 2017 the Australian Council for Educational Research (ACER) in co-operation with the Afghanistan Ministry of Education undertook a program of national assessment in Afghanistan: the Monitoring Trends in Educational Growth (MTEG). The assessment program measures mathematical, reading and writing literacy in the national population at the end of grade 3 and grade 6. An aim that emerged later in the assessment program was to be able to locate grade 3 students and grade 6 students on the same learning metric and to describe growth not only within the same grade cohort, but between them in the long-term. This study comprises of, in grade 6, 5 979 students, in grade 3, 4 936 students, and in a grade 4 and 5 link sample, 1200 students.

This paper addresses the methodological approaches used to meet this challenge. The measurement approach had to control for variations in modality - Grade 6 was a paper- based assessment while grade 3 was computer-based to accommodate the relatively low levels of literacy in the grade 3 population. This linking study also had to implement a novel design, with neither common-students or items, an intermediate sample of grade 4 and 5 students was drawn. Novel approaches to assessing the quality of link-items was needed, and both model-oriented approaches (e.g., iterative comparisons of nested model deviance) sensitivity analysis (e.g., rank-order association of item parameters between samples) were implemented with the relatively sparse data.

This study illustrates how novel approaches are needed in fast-moving development contexts. This study will also demonstrate how the work being implemented right now can be used to support capacity development and growth within countries and to aid increased engagement with the international community through the SDG 4 agenda. Demonstrating that countries can report against the SDGs while using their own assessment programs is an important step to increasing the engagement of all countries in the learning for all agenda.

Creating a vertical scale to support the Sustainable Development Goal agenda of lifelong learning

Claire Scoular, Dan Cloney, Alex Daraganov, Ray Adams, Ross Turner, Leigh Patterson

Session 5A, 13:00 - 14:30, HAGEN 2

This paper presents an example of how educational measurement can contribute to the next generation of assessment systems. It outlines a joint initiative by Australian Council for Educational Research Centre for Global Education Monitoring (ACER-GEM) and UNESCO Institute of Statistics (UIS) to develop empirically supported vertical scales in mathematics and reading. The scales will play a role in improving the quality of measuring and monitoring learning outcomes within countries (including those in developing and conflict-affected contexts), and address the challenges associated with between-country comparisons. Such advancement is essential to ensuring the Sustainable Development Goal (SDG) agenda is achieved and that all countries, including those not participating in large scale assessments, have the opportunity to participate and benefit.

The vertical scales describe learning progressions for reading and mathematics, across the range of proficiencies that typically develop throughout compulsory schooling. The aim is to enable countries to examine and report the outcomes of their assessment activities using a common methodology. Despite the high level of participation in learning assessments, clearly defined vertical scales and intra- as well as inter-assessment comparability remain limited. This presents particular challenges for measuring progress against the SDGs for learning outcomes. The learning goals and targets will only have meaning and utility if they are underpinned by empirically derived

common scales that accommodate results from a range of different assessment programs. Vertical scales provide a means to assess the emerging competencies of learners, and to explore cognitive growth and trends in growth over time. The development of the vertical scales allows policy makers, education practitioners and education investors to not only quantify and compare learner proficiency, but also describe it in a meaningful way.

The vertical scales are based on an empirically analysis of the relative difficulties of items across assessment programs international, grounded in a conceptual framework taking in the current state-of-the-art of reading and mathematics theory. To permit comparison of the difficulty of the different item sets mapped to the vertical scales, a pairwise comparison methodology (BTL model) was employed. More than 500 items from 14 assessment programs were included in the analysis and more the 30 000 comparisons were made for each of reading and mathematics . The purpose of this comparison was to generate a set of difficulty estimates across the entire item set used in the initial steps of development of the vertical scales for reading and mathematics respectively. A pairwise comparison of items enables the different assessment programs from which those items were sourced to be aligned, allowing inferences to be made as to the underlying learning progression represented by the items. By modelling the cumulative information provided by multiple comparisons from many content specialists, estimates of the difficulties of items on a latent scale were obtained. Excellent evidence was generated that the pairwise estimates recovered the within-assessment program item parameters (where published). This presentation will present the methodology undertaken to create the vertical scale, outline the broad findings, and discuss implications and next steps for validation.

Large-scale alternate assessments based on fine-grained learning maps: Opportunities and challenges

Meagan Karvonen

Session 5A, 13:00 - 14:30, HAGEN 2

In the United States of America, students with the most significant cognitive disabilities participate in statewide academic assessment systems through alternate assessments based on alternate achievement standards (AA-AAS). The population of students who take AA-AAS is very small (approximately 1% of students) and extremely heterogeneous (Kearns et al., 2011). AA-AAS were first conceived nearly 20 years ago. Since then the educational assessment field has dealt with tensions between the standardization typical of large-scale assessment and the flexibility needed to ensure accessibility for the population. Due to design constraints and the population, AA-AAS also have unique challenges with regard to evidence of validity and technical quality.

In 2010 a consortium of states began developing a next generation AA-AAS. First used operationally in 2015, the Dynamic Learning Maps (DLM) alternate assessment system now serves 90,000 students across 18 states. DLM assessments are based on large, fine-grained learning maps with thousands of nodes (skills) and multiple pathways by which students develop understanding of academic domains (Kingston, Karvonen, Bechard, & Erickson, 2016).

Assessments are designed using a combination of evidence-centered design and universal design principles. Assessments are delivered in short testlets with varying degrees of complexity relative to the content standard. Unlike most large-scale academic assessments, the DLM system goes beyond summative uses. Testlets are designed to be instructionally relevant. Teachers select and use instructionally embedded assessments throughout the year so results guide instruction. Consistent with the highly multidimensional nature of the learning maps, DLM assessments are scored using Cognitive Diagnostic Modeling. Summative results are based on aggregated mastery of discrete skills. Score reports feature fine-grained, diagnostic information to guide instruction as well as summative results used for program evaluation and accountability.

The proposed session for the innovative assessment strand will begin with a brief description of the philosophical underpinnings and design of the DLM alternate assessment system. Several opportunities and challenges will be described in more depth, using evidence from early development and four years of operational test administration. Depending on the length of the session, topics would likely include: (1) evaluation of the test development approach that integrates evidence-centered design and universal design for learning; (2) implementation evidence interpreted in light of the program's theory of action; (3) an overview of the modeling research needed to support the use of CDM for scoring; and (4) the standard setting approach designed for use with CDM-based results. The session concludes with a summary of future directions for the DLM system and potential implications for other assessment systems.

Performance assessment of learning in higher education (PAL)

Richard J. Shavelson, Olga Zlatkin-Troitschanskaia, Susanne Schmidt, & Klaus Beck

Session 5A, 13:00 - 14:30, HAGEN 2

The demand to measure higher education outcomes has gained worldwide momentum. While there are many approaches to measuring higher education learning outcomes, including self-reports of learning and multiple-choice tests, the PAL study's focus lies on performance assessment of learning with particular focus on the measurement of so-called 21st century (generic) skills such as critical thinking (CT) and critical reasoning (CR), with a task that simulates real-life decision making and judgment situations (e.g., Shavelson, 2013; Shavelson et al., 2015, 2018).

The assessment framework is based on a performance task (PT) that demands CT and CR and resembles the myriad of complex everyday life situations. A real-world event is presented along with information more or less relevant to the problem. The problem requires CT and CR in terms of recognizing and evaluating the relevance, reliability and validity of the given information as well as evaluating the problem and finally making a decision. Information regarding decision making and thought processes, particularly regarding the ability to deal with a huge amount of partly irrelevant and unreliable information was gathered in a semi-structured cognitive interview after the completion of the PT (N=30 undergraduate students).

The PT is delivered on a computer and the information needed to solve the problem is presented within the task itself as well as in full length over the internet (such as newspaper or Wikipedia articles). Computers provide substantial leeway both in delivering tasks and in their fidelity to the real world they are intended to emulate. The format is open-ended, students constructed answers of varying length in response to the prompt inviting them to make a decision about the real-world event. The difficulty of the task is fine-tuned through the way the information is presented, the number of information sources and points to consider, including distractors (irrelevant information), and their trustworthiness and relative strength compared to one another as well as time constraints and response requirements.

For the response ratings, analytic categories were developed based on the construct definition of CT and CR (Shavelson et al., 2018). These categories consider the students' use of (un)reliable and (in)valid information as well as their reflection and avoidance of heuristics that lead to errors in judgment and decision making. The students' use of such information for justifying decisions, problem solving and/or recommendations for action are evaluated. Moreover, argumentation, the use of evidence to support claims, and clarity of communication is rated. Additional aspects were revealed within the cognitive interviews, which goes hand in hand with the analyses using mixed-methods. Based on the coding scheme of the interview (which is in line with the "Grounded Theory"), we quantified the codes in accordance to the construct definition of CT and CR. By doing so, analyses indicated that, for instance, many students knew that Wikipedia is no trustworthy reference but most of them used it for their argumentation within their statement. These and further cognitive processes are modeled within a multilevel mixed model following the approach by Brückner and Pellegrino (2017).

Asymmetry in fixed-precision M-CAT: Multidimensional selection versus marginal stopping

Johan Braeken & Muirne C. S. Paap

Session 5B, 13:00 - 14:30, VIA

Standard implementations of a Multidimensional Computerized Adaptive Testing (M-CAT) algorithm have item selection rules that are searching for items that optimize the Fisher information volume. A variable-length M-CAT would usually include a stopping rule requiring all dimensions being measured with a fixed minimum precision. In contrast to the inherently multidimensional selection rule, this stopping rule is defined at the marginal levels of the latent traits distribution: standard error smaller than a pre-determined threshold value for each dimension. This asymmetry between selection rule and stopping rule leads to side-effects that might not always be anticipated at first glance. We will first revisit and discuss the issue from a distribution and practical perspective, subsequently propose some work-arounds in the form of alternative selection rules, and elaborate on their effectivity to tackle the issue in practice.

Comparing multidimensional to unidimensional computerized adaptive testing under two empirical scenarios: The impact of design factors

Muirne Paap, Sebastian Born, & Johan Braeken

Session 5B, 13:00 - 14:30, VIA

Research has shown the benefits of taking into account the correlation among dimensions when estimating latent trait scores in computerized adaptive tests (CATs). Multidimensional CATs (MCATs) could further improve measurement precision/decrease test length as compared to using separate unidimensional CATs for each domain, especially if domains are highly correlated.

In this study, we systematically evaluate the impact of a number of important design factors on CAT performance, using realistic example item banks. Two main scenarios are compared: health assessment (polytomous items, small to medium item bank sizes, high discrimination parameters) and educational testing (dichotomous items, large item banks, small to medium-sized discrimination parameters). Measurement efficiency is evaluated for both between-item multidimensional CATs (MCAT conditions) and separate unidimensional CATs for each latent dimension (UCAT condition). We focus on fixed-precision CATs since it is both feasible and desirable in health settings; but to date most research regarding CAT has focused on fixed-length testing. This study shows that the benefits associated with fixed-precision multidimensional CAT hold under a wide variety of circumstances.

MCAT has great potential when it comes to reducing test length and improving accuracy and precision of latent trait scores, both in health and educational measurement. We will discuss how the incremental value of MCAT depends on factors like adequate targeting, the size of the correlations, item bank size, and item parameters.

Stochastic programming for automated test assembly with uncertainty in the item parameters or in the responses

Bernard Veldkamp

Session 5B, 13:00 - 14:30, VIA

Items can be described by many parameters. They can be related to, for example, the content of the item, the psychometric properties, or the process of solving the item. Nowadays, response time parameters receive a lot of attention. Early research on response time modeling assumed that a test taker would show consistent response time behavior, often referred to as working speed, over the course of a test. Such models were unrealistic for various reasons — a warm-up effect may cause a test taker to respond more slowly than expected to the early items, fatigue may cause a test taker to respond more slowly than expected toward the end of a test, or as time runs out the test taker may quickly guess the answers to the last items on a test. To take these variations in working speed into account, mixture response time models have recently been investigated. When these models are applied in automated test assembly, probabilistic response

time constraints have to be imposed. Stochastic programming has been applied to deal with this kind of probabilistic constraints. In the current paper, the application of stochastic programming will be generalized to uncertainties in the model, for example coming from automated item generation or open answer questions.

Continuous item calibration in computerized adaptive testing

Andreas Frey, Aron Fink, Sebastian Born, & Christian Spoden

Session 5B, 13:00 - 14:30, VIA

In computerized adaptive testing (CAT), knowledge about the item parameters of the test items in the pool is required to select the next item. These item parameters are estimated based on item responses collected in a calibration study using an item response theory (IRT) model. In several potential application areas of computerized adaptive testing (CAT), constructing large numbers of items prior to the test's initial use, and/or carrying out a calibration study with a large sample is not feasible. Correspondingly, for applications such as written standardized exams, psychological tests used in personnel selection, for clinical diagnoses or in research, CAT is typically not used, even though it would be advantageous here too. To extend the application range of CAT, a new continuous calibration strategy is presented and illustrated. This calibration strategy is applicable when setting up a CAT anew or when converting a linear test into a computerized adaptive test. The basic ideas of the strategy are (a) item calibration oriented on the time and capacity available for test development, (b) utilizing item responses across periodical assessments for item calibration purposes, (c) maintaining the measured scale over time, and (d) continuously increasing the adaptivity of the test during its operational use. In the presentation, I will describe the key elements of the new continuous calibration strategy and present results from a comprehensive simulation study. The simulation is based on a factorial design with the between factors IRT model (1PL, 2PL), sample size (50, 100, 300), item parameter estimation method (MML, Bayesian), and the within factor test cycle (1, ..., 11). The results showed a promising performance of the proposed strategy even for very small sample sizes. Based on a detailed presentation of the results, I will conclude with aspects of the continuous calibration strategy that should be covered by future research prior to its operational use.