Frontier Research in Educational Measurement (FREMO)

SEPTEMBER 5-7, 2023

Conference Abstracts
# Table of Contents

WORKSHOP: EQUATING TEST SCORES WITH R ................................................................. 2

KEYNOTES ..................................................................................................................... 3

PARALLEL SESSIONS .................................................................................................... 7

  Session 1A: “Estimation and model fit” ..................................................................... 7
  Session 1B: “Measurement of non-cognitive constructs” ........................................... 12
  Session 2A: “Equating, linking and scaling” ............................................................. 17
  Session 2B: “Modern test design” ............................................................................ 21
  Session 3A: “Multilevel modeling” ........................................................................... 26
  Session 3B: “Automated scoring and item generation” ............................................. 30
  Session 4A: “Assessment designs” .......................................................................... 33
  Session 4B: “Disengaged responding” ..................................................................... 38
  Session 5A: “Adaptive testing” ................................................................................. 42
  Session 5B: “International large scale assessments” ................................................. 47

POSTERS ....................................................................................................................... 51
Workshop: Equating test scores with R

Abstract
The aim of test score equating is to adjust the test scores on different test forms so that they can be comparable and used interchangeably. The overall goals of the workshop are to help the attendees understand the principles of equating, to conduct equating, and to interpret the results of equating. Different R packages will be used to illustrate how to perform equating when test scores data are collected using different data collection designs. Traditional equating methods, item response theory (IRT) equating methods and kernel equating methods are included in the workshop.

Presenter
Prof. Marie Wiberg is a Professor at Umeå School of Business, Economics and Statistics (USBE). Her research interests include educational measurement and psychometrics in general, especially test equating, parametric and nonparametric item response theory, and international large-scale assessments.
Equating test scores across diverse groups poses a significant challenge in educational and psychological assessment. During times of societal crisis, significant changes unfold, affecting various aspects of society. Admissions tests, intended for all individuals, may witness alterations in the composition of test taker groups. This composition exerts a profound impact on the test equating process. When substantial changes occur, it becomes uncertain whether groups taking different test forms can be assumed equivalent. This presentation delves into the issue of equating test scores when confronted with nonequivalent groups. Through the exploration of innovative approaches, we examine the complexities and potential solutions for attaining fair and precise test score equating across diverse populations. The discussion encompasses the influence of nonequivalent groups on test score equating and offers strategies to ensure equitable assessment practices.
Interrater Reliability and Optimal Weights for Components of Multidimensional Observational Instruments

Cees A.W. Glas

University of Twente, the Netherlands

Abstract

In educational research, part of the data may consist of observations by multiple raters at different tasks or time points collected using itemized observation instruments. Further, the data often have a multilevel structure pertaining to students nested within teachers or classes. The model for analyzing such data can involve the combination of an item response theory (IRT) model for the item responses, a generalizability theory (GT) model for different raters at different time points, and a multilevel (ML) model to connect outcome variables with predictors.

However, in some instances, the observed construct (say teacher proficiency) may not be regarded as unidimensional on theoretical grounds, or a one-dimensional representation of the construct may not fit the empirical data. In such cases, a multidimensional IRT model can be used to model the item responses pertaining to the different subscales.

Still, in much research and many applications, there is a need to combine the latent subscale scores into an overall index.

A method is proposed for constructing such indices as linear functions of variables such that the reliability of the index is maximized. Reliability is defined in the framework of a combination of an IRT model and a GT model and optimal weights of the components of the index are found by maximizing systematic variance relative to the total variance in the measurements. The model parameters, including reliability indices and optimal weights, are estimated in a Bayesian framework using Markov chain Monte Carlo (MCMC) computational methods.

Several examples from ongoing educational research projects are presented.

Keywords

Bayesian estimation, full-information factor analysis, generalizability theory, item response theory (IRT), Markov chain Monte Carlo, multidimensional item response theory (MIRT), optimal weights.
The mixture approach to finding measurement invariance across countries
Kim De Roover
KU Leuven

Abstract
In educational measurement – for instance, in international assessment programs like PISA and TIMMS – student’s achievements are often related to non-cognitive constructs such as their academic self-concept and self-efficacy, their enjoyment of science, sense of belonging, and wellbeing. A critical assumption in cross-national comparative research is that the construct(s) are measured in exactly the same way across all countries (i.e., measurement invariance). The comparability of the non-cognitive constructs is questionable and rarely evaluated, though. When measurement invariance does not hold, one may compare country-specific measurement models to pinpoint non-invariances. But the number of countries involved in the international student assessments is very large which makes the pairwise comparisons infeasible and anything but insightful. To make it easier to unravel invariances from non-invariances and for which countries they apply, an intuitive solution is clustering the countries into a few clusters based on the measurement model parameters by means of a mixture model. Not only does this pin down the number of comparisons needed to identify non-invariances, but the clustering of the countries is an interesting result in itself and provides clues on how to move forward with the cross-national comparisons or further measurement invariance testing. To this aim, I recently presented mixture multigroup factor analysis (MMG-FA; De Roover, 2021; De Roover, Vermunt, & Ceulemans, 2022) to clusters groups (e.g., countries) according to a specific level of ‘clusterwise’ measurement invariance (e.g., based on factor loadings only to achieve metric invariance within clusters, or based on loadings and intercepts to achieve scalar invariance within clusters). MMG-FA accommodates a unique blend of cluster- and group-specific parameters to set aside parameter differences that are irrelevant to the measurement invariance question and to assure the clustering focuses on differences in measurement. In this way, it outperforms other mixture factor analysis methods.
Potential benefits and challenges of log data in large-scale assessment
Frank Goldhammer
Leibniz Institute for Research and Information in Education, DIPF

Abstract
International large-scale assessments such as PISA, PIAAC, and TIMSS have begun to provide log file data, i.e., events, event-related attributes, and timestamps of test takers' interactions with the assessment system. The log data and the process indicators derived from them can be understood as behavioral correlates of the test taker's response process. In this presentation, the role of individual differences in response processes will be discussed from a measurement perspective. Based on that, the potential benefits of using log data for educational measurement as well as related challenges will be addressed. For example, log data enable the measurement of new process-related constructs, they can enhance scoring, increase measurement precision, or help improve data quality post hoc or in situ. The interpretation of process indicators requires validation to make sure that inferences, for instance, in terms of cognitive or motivational states, are justifiable both theoretically and empirically. Another challenge is the dissemination of log data collected in large-scale assessments, their accessibility, and documentation for reproducible analyses.
Parallel sessions

Day 1, Wednesday, September 6, 2023

Session 1A: “Estimation and model fit”
Chair: Cees Glas
Room: Hagen 1-2

Estimating Interrater Reliability for Planned Missing Observational Designs
Debby ten Hove, Letty Koopman

In observational research, raters are used to gather information about attributes of subjects. For research purposes, it is important that the observed attribute scores reflect the attributes of the subjects, rather than differences across raters and their perspective. Therefore, it is desired to estimate the interrater reliability (IRR) of the observations. Many observational studies use a planned-missing observational design to distribute the workload across raters. This is especially the case in large-scale observational assessments, in which constraints due to logistics or availability of individual raters dictate the choice for a planned missing observational design. This, in turn, results in incomplete data. In incomplete designs, the raters partly vary across subjects. Intraclass correlation coefficients (ICCs) are very flexible IRR estimators that serve many research purposes. ICCs can for example estimate both agreement and consistency across both single and averaged ratings. However, traditional analysis of variance based estimation methods for the ICC cannot compute the IRR from incomplete data, as they require the raters to be either unique or identical for each subject. Recent developments allow the estimation of ICCs from incomplete data, and recent simulation studies showed that a maximum likelihood estimation (MLE) approach of random effects models is the preferred choice. In this presentation, we introduce a newly developed Shiny application that can be used to apply this MLE method to estimate the IRR from incomplete data. We will pay special attention to the various choices that researchers have to make among the different ICCs that can be computed with the tool. Moreover, we will discuss the implications of design choices on the magnitude of the ICC and the precision of its estimates.
Longitudinal Modeling of Age-Dependent Latent Traits with Generalized Additive Latent and Mixed Models
Øystein Sørensen

I present generalized additive latent and mixed models (GALAMMs) for analysis of clustered data with responses and latent variables depending smoothly on observed variables. A scalable maximum likelihood estimation algorithm is proposed, utilizing the Laplace approximation, sparse matrix computation, and automatic differentiation. Mixed response types, heteroscedasticity, and crossed random effects are naturally incorporated into the framework. The models developed were motivated by applications in cognitive neuroscience, and two case studies are presented. First, we show how GALAMMs can jointly model the complex lifespan trajectories of episodic memory, working memory, and speed/executive function, measured by the California Verbal Learning Test (CVLT), digit span tests, and Stroop tests, respectively. Next, we study the effect of socioeconomic status on brain structure, using data on education and income together with hippocampal volumes estimated by magnetic resonance imaging. By combining semiparametric estimation with latent variable modeling, GALAMMs allow a more realistic representation of how brain and cognition vary across the lifespan, while simultaneously estimating latent traits from measured items. Simulation experiments suggest that model estimates are accurate even with moderate sample sizes.
Fast estimation of longitudinal generalized linear latent and mixed models with a second-order Laplace approximation

Björn Andersson

Maximum likelihood estimation of high-dimensional latent variable models is computationally demanding due to the need to approximate the intractable integrals in the likelihood function. We introduce a fast approximate maximum likelihood estimation method for generalized linear latent and mixed models, based on second-order Laplace approximations of the integrals in the likelihood function. Estimation is implemented with a quasi-Newton method that uses the analytical gradient of the Laplace-approximated log-likelihood function together with the BFGS approximation of the Hessian matrix. We apply the method to repeated measurements of a mixture of categorical, count and continuous observed variables and explain how the computational efficiency of the implementation depends on the complexity of the model structure. The graded response model and the generalized partial credit model are supported for categorical data, the Poisson and negative-binomial distributions are supported for count data, and the normal distribution is supported for continuous data. The method is compared against alternative approaches in simulation studies with non-linear latent growth models and two-tier longitudinal models, demonstrating high computational efficiency and accurate parameter recovery. We discuss statistical inference using likelihood ratio testing and give an overview of methods for estimating standard errors via different types of approximations of the observed information matrix and the sandwich estimator. An R implementation of the method, lamle, is presented and some remaining estimation challenges are outlined.
Non-parametric regression among factor scores
Julien Irmer, Steffen Grønneberg

Non-linear structural equation modeling (NLSEM) starts with the specification of a parametric structural regression function. In contrast to standard regression problems with directly observed variables, we cannot easily use graphical and non-parametric methods to motivate the parametric form of the regression function. We provide reasonable assumptions that identify this regression function non-parametrically, and show that under these conditions, we can input Bartlett factor scores into a class of non-parametric regression procedures that take into account measurement error to get consistent estimates of the structural regression function. This allows the psychometrician to motivate and error check parametric regression assumptions in NLSEM.
An evaluation of factor model fit procedures under non-normality and high dimensions
Njål Foldnes

We present results from a large-scale Monte Carlo study of fit statistics for factor models in high dimensions, where state-of-the art non-normal simulation methods are employed. We include the understudied Bollen-Stine bootstrap test, as well as a newly developed set of test procedures based on eigenvalue block averaging. Also, a bootstrap-based selection procedure is evaluated. Our results indicate that non-normality may pose a serious threat to well-known and widely used statistics such as the Satorra-Bentler and the scaled-and-shifted test. Overall, eigenvalue block averaging statistics outperform these statistics.
The Relationship between Academic Resilience and Psychological Well-Being – an Analysis of Swedish PISA 2018 data
Deborah Elin Siebecke, Kajsa Yang Hansen, Maria Jarl

Research suggests that students from socioeconomically disadvantaged backgrounds, in general, demonstrate lower academic achievement (e.g., Sirin, 2005). However, resilience theory is grounded in the recognition that individuals’ responses to adversity differ (Rutter, 2012) and some students demonstrate positive adaptation despite adversity. The present study focuses on academically resilient students in Sweden, defined as those who “beat the odds” and achieve high despite their socioeconomic disadvantage. The study's main objective is to investigate the relationship between academic resilience and psychological well-being.

Previous research found that academically resilient students reported higher life satisfaction compared to non-resilient students (OECD, 2019). Life satisfaction can be regarded as part of students’ psychological well-being, yet more research regarding the complex construct of psychological well-being is needed. Making use of Swedish data from the Programme for International Student Assessment (PISA) from 2018, the present study operationalizes academically resilient students as those who fall within the bottom quartile of the Swedish distribution of the Index of Economic, Social, and Cultural Status and achieve at or above Level 3 in all three PISA domains mathematics, science, and reading – a level that is said to prepare students “for success later in life” (Agasisti et al., 2018, p. 8). Psychological well-being is measured using indicators of life satisfaction, eudaimonia (i.e., students’ meaning and purpose in life), and hedonia (students’ experience of certain positive and negative emotions) (Borgonovi, 2020). Analyses were carried out in SPSS 29 and Mplus 8. Firstly, confirmatory factor analysis was used to test the measurement property for the well-being constructs eudaimonia and hedonia and resulted in an overall good model fit. Secondly, structural models were built to connect the psychological well-being dimensions with the endogenous resilience variable. Results indicate that psychological well-being, to a certain extent, relates to academic resilience. Yet, the relationship is not straightforward and requires critical discussion.
Examining psychometric properties of Woodworth Psychoneurotic Inventory under binary IRT models
Hsin Kao, Chia-Wen Chen

For the first English-published self-report personality measurement, Woodworth Psychoneurotic Inventory (WPI) was developed to test the emotional instability of army servicemen throughout World War I. Although the military applications of WPI were short-lived, it had influenced various personality scales developed over the following 30 years. The importance of WPI has been addressed in the history of personality measurement. However, a lack of studies investigated the psychometric properties and dimensionality of WPI under the item response theory (IRT) framework. We address this research gap by applying IRT models to a large-scale dataset with 6,019 examinees across 110 countries participating in a WPI. We conducted two studies to examine the psychometric properties and dimensionality of WPI. For the first study, since WPI was originally assumed to have only one dimension, we fitted a unidimensional IRT model and evaluated item fit indices and G2 statistics for the analysis of local dependence under the unidimensional assumption. The result of study 1 showed that more than half items exhibited a significant lack of item fit and identified as local dependence, which may indicate that WPI violates the unidimensional assumption. The second study examined the dimensionality of WPI by using the dimension detection method called DETECT. The result suggested that WPI consists of four dimensions and better fits a multidimensional 2PL model than a unidimensional model in AIC and BIC indices. Although some pairs of items were still detected to have local dependencies after controlling for the latent variables, the number markedly decreased in fitting a four-dimensional model compared to a unidimensional model. In addition, most of the items and persons fit were also improved. In conclusion, we suggested that the four-dimensional 2PL model is more advisable than the unidimensional one for the WPI.
Validity and comparability of the leisure time questionnaires in the Evaluation Through Follow-up study
Erika Majoros, Thea Klapp, Alli Klapp

The Evaluation Through Follow-up (Utvärdering Genom Uppföljning, UGU) study is a large cohort-sequential database that is used for evaluation and research about schools and education in Sweden (Härnqvist, 2000). Administrative data, questionnaires, various measures of students' learning conditions, and educational outcomes have been collected at several time points during school years for ten cohorts since 1961. Some of the instruments have been revised over time, which might introduce unintended comparability issues. However, little research has been looking into these instruments concerning measurement invariance within and across cohorts. Therefore, there is a need to evaluate the influence of questionnaire revisions at the scale and item levels. Factorial measurement equivalence or invariance refers to whether scale scores of an assessment have the same meaning under different conditions, e.g., administration methods, countries, or populations (Kline, 2016). Lack of item bias refers to the invariance of the item parameters across groups of interest (Meredith & Teresi, 2006). When this invariance is not fulfilled, differential item functioning (DIF) occurs. The present study aims at investigating the measurement of leisure activities for the cohorts born in 1998 and 2004 both at the scale level and the item level. Methods in the structural equation modeling and classical and modern test theory frameworks are applied. The findings of this study inform the questionnaire development in the UGU project as well as secondary analyses of this unique longitudinal database. References Härnqvist, K. (2000). Evaluation through follow-up. A longitudinal program for studying education and career development. In Seven Swedish longitudinal studies in behavioral sciences (pp. 76–114). Swedish Council for Planning and Coordination of Research Forskningsrådsnämnden FRN.
The purpose of the study is to investigate the comparability of latent constructs measuring academic self-concept and self-efficacy (ASC and ASE) across and within student groups. To make meaningful comparisons of latent measurements across cohorts and time points, the measurements need to be equivalent across groups, i.e., representing the same latent phenomenon. Established measurements are commonly used to operationalize self-perceived abilities (e.g., the Self-Description Questionnaire; Marsh, 2007). As it may not be possible to use established measurements with secondary data, examples of research designs that test the comparability of alternative measurements are needed. Data from the Swedish longitudinal infrastructure Evaluation through Follow-up (UGU) will be analysed. UGU contains questionnaire data from multiple birth cohorts and time points. Latent variables measuring ASC and ASE will be compared across the birth cohorts 1992, 1998, and 2004 and longitudinally within each cohort. The study aims to provide an example of how ASC and ASE measures may be tested for measurement invariance with multiple-group confirmatory factor analysis (CFA). Configural, metric, and scalar invariance will be considered. If metric invariance is supported, it is possible to compare unstandardized relations involving the latent variable across the groups. If the most restrictive level of invariance is supported, i.e., scalar invariance, it is possible to compare latent means across the groups (Leitgöb et al., 2023).
Motivational beliefs and writing performance in Peruvian secondary school students: a multigroup analysis.
Frank Villegas

Motivational beliefs play an important role in the writing process. Some beliefs stand out for their predictive role in writing performance: self-efficacy, achievement goals, and beliefs about the malleability of writing. The objective of the research is to analyze the structural relationship between motivational beliefs (self-efficacy, beliefs about the malleability of writing, and achievement goals) and writing performance across urban and rural groups. To do this, the responses of 5968 second-grade high school students in Peru were considered, who answered a writing test and scales in the 2018 Sample Assessment, a large scale assessment implemented by the Peruvian Ministry of Education. A hypothetical model was evaluated using structural equation modeling and a multigroup analysis. These analyses considered the characteristics of the complex sample design (sample weights, clusters, and strata). The results show that a greater incremental approach is associated with a greater establishment of domain goals, and that a greater fixed approach is associated with a greater establishment of performance goals. In addition, a greater establishment of domain goals of approach is associated with better writing performance, while a greater establishment of performance goals (approach and avoidance) is associated with lower writing performance. These relationships are similar in the multigroup analysis, although a greater magnitude is found in the rural area in the predictive role of approach domain goals and avoidance performance goals in writing performance. The results suggest the importance of promoting approach domain goals and an incremental approach to writing in writing education. They also highlight the need to continue efforts to implement the communicative approach in the Peruvian context.
Improved item response theory ability scales using information theory
Joakim Wallmark

Item response theory (IRT) is a statistical method that assesses the measurement properties of test items and estimates test taker abilities through the analysis of their responses on the items. However, with traditional IRT, ability estimates are usually put on an arbitrary scale that lacks meaning in terms of distance. In this study, we demonstrate how information entropy can be used to create IRT scales with metric properties. We also show how entropy scales can serve as a replacement for sum scores to measure the amount of information contained within a test obtained by a test taker. Additionally, we illustrate how these scales can be used to bring scores from various IRT models to a common scale.
Discrete kernels for smoothing score distributions
Jorge González, Marie Wiberg

Although the sample relative frequency distribution can be used as an estimate of a score distribution, presmoothing is often used to estimate and correct irregularities of score distributions. Both, log-linear and mixture models (i.e., Beta4) have been used for presmoothing in equating. In this paper, we introduce a class of discrete kernels that can be used to estimate the probability mass function of scores, serving thus for the purpose of presmoothing. An empirical illustration shows that the proposed discrete kernel estimates work either equally well or better than current methods for presmoothing score distributions.
Obtaining comparable scores from multiple test forms in case of non-equivalent groups via repeated covariate equating
Patrícia Martinková, Michaela Vařejková, Eva Potužníková

The traditional way to equate scores from alternate test forms is to include anchor items which are administered to all test takers and are used to adjust for possible differences in ability between the groups taking different forms (Kolen & Brennan, 2004, von Davier, 2011, von Davier, 2013, González & Wiberg, 2017). In the absence of anchor items, one approach is the so-called covariate equating, in which the anchor items are substituted by covariates, such as grades or other test scores (Wiberg & Bränberg, 2015, Longford, 2015, Wallin & Wiberg 2019). This procedure assumes that the covariates can account for differences in ability and that the conditional distribution of the test scores, given the covariates, is the same for both groups. In this work, we consider different violations of the assumptions of the covariate equating method, motivated by real data examples from Czech matura examinations. We consider the case in which the correlation between the covariate and the test score is weak, thus the assumption that the covariates explain the difference in ability is not met. We also consider the case in which the covariate itself is measured using different test forms, thus violating the assumption of the same conditional distribution. We conduct a simulation study showing that, for the latter case, equating the covariate before incorporating it into the primary test scores equating algorithm can improve the accuracy of the resulting equated scores. The real data example demonstrates the proposed repeated covariate method, as well as the effect of low correlation between the score and the covariate.
Streamlining Item Generation: A Visual and Numerical Systematization of Qualitative Data from Multiple Sources
Tine Nielsen, Morten Pettersson

Developing a non-ability scale for an education-relevant construct is a challenging task. Typically, researchers begin with a qualitative approach, such as conducting interviews with multiple stakeholders and gathering information from other sources. This often generates a large amount of data that is difficult to analyze and organize in a systematic way, which facilitates item generation. To address this issue, we propose a method that involves the following steps: 1) Extract construct-relevant information (statements) from qualitative sources. 2) Conduct a content analysis to identify overlapping statements and avoid redundancy by combining them into unique statements. 3) Evaluate the level of agreement on unique statements across different stakeholders using a matrix of statements and stakeholders (1 = mentioned by a stakeholder, 0 = not mentioned). For a visual evaluation, we use an expanded-for-purpose version of Fried’s (2017) co-occurrence circle, which provides a single graphical overview of the agreement across 11 stakeholders and objectives. For a numerical evaluation, we calculate Jaccard’s similarity coefficient for all pairs of stakeholders using the formula $s/(u_1 + u_2 + s)$, where $s$ represents the number of common statements between two stakeholders, and $u_1$ and $u_2$ represent the number of statements unique to each stakeholder.

To illustrate this proposed method, we use an ongoing study as an example. This study focuses on developing a new scale (FPSES) that measures field practice self-efficacy among student teachers. In order to identify which teaching practices student teachers may have difficulty believing they can enact or engage in while in field practice, the study collects qualitative data from a plethora of sources. These include two focus-group interviews with student teachers, four single interviews with field practice teachers, four written accounts from campus teachers, one written account from a field practice coordinator, and official documents outlining the field practice objectives.
Semi-adaptive designs: The case of the Norwegian numeracy mapping test for primary education
Guri A. Nortvedt, Henrik H. Haram, Karianne Berg Bratting, Oksana Kovpanets, Andreas Pettersen

The aim of this presentation is to present a semi-adaptive test design used in the Norwegian national numeracy mapping tests. The mapping tests are developed for grades 1 and 3 students and they use a two-step assessment design to identify students in need of remedial teaching. In the two steps, students are filtered out if “the test” deems them to not be in need of remedial teaching: in step 1, approximately 70% of the students are filtered out, and in step 2 another 10% of students are filtered out. The remaining 20% of students are likely in need of remedial teaching and are identified with sufficiently high precision. The tests and the items within it are developed using IRT-analyses to select the best fitting items for each of the two test-steps, however, the test delivery platform does not include an IRT module. Consequently, the cut-scores have to be converted from theta-scores into points. Such conversions come at the risk of large differences in which students are identified as in need of remedial teaching. Applying data from the first implementations of the Norwegian numeracy mapping tests, we will discuss risk of identifying the “wrong” students, i.e. consequences for type I and II errors.
Optimal item calibration for the Swedish national test in Mathematics - design, IRT analysis and results
Ellinor Fackle Fornius, Frank Miller

For large achievement tests, such as the national tests in school, item calibration is used to determine the characteristics of an item, which are important to learn before the item is used in a operational test. There exist methods for optimal calibration that allocate items to examinees based on examinee ability, when examinees arrive at the test sequentially. For the Swedish national test in Mathematics, the item calibration is instead conducted in parallel. In addition, the methods often assume a single item format, such as dichotomous items, while the national tests typically consist of items of mixed format, including also e.g. polytomous items and multiple choice items. A recently developed method for optimal calibration that handles parallel calibration as well as mixed format tests was evaluated in a calibration study conducted in Swedish schools during the spring of 2022. In this talk, we describe the optimal calibration design derived for this setting. We illustrate for which item types the optimal design performs best (in terms of efficiency) and how expected time needed per item was taken into account. We also highlight relevant aspects of the IRT analysis and demonstrate some results and conclusions of the study.
Computerized Adaptive Testing (CAT) is a promising tool for reducing the length of educational assessments. However, when it comes to high-stakes tests, additional considerations are needed to ensure test security, such as item exposure control. The Progressive-Restricted method (PR) has been used to address this issue by limiting item exposure without compromising test precision. The PR method uses a random component that reduces in importance as the test progresses, and the acceleration parameter determines the speed of this reduction. However, little research has been done on this parameter, especially in variable-length CATs. Therefore, this study aims to evaluate the PR method with different acceleration parameters in both fixed and variable length CATs, using the Brazilian National High School Exam (Enem) as a case study. We combined two item selection methods, Maximum Fisher Information (MFI) and PR, with three stopping rules and simulated CATs for the four Enem tests. We compared test length and precision for each condition and analyzed the impact of the acceleration parameter on item exposure. Additionally, we compared the linear format of the Enem with a 20-item CAT.

Our results showed that test length was consistently lower with MFI than with PR. In fixed-length CATs, MFI and PR had similar precision with the standard error rule, but PR had lower precision with the error reduction rule. Furthermore, item exposure decreased as the acceleration parameter of the PR increased, which improved test security. The 20-item CAT had greater accuracy than the linear format of the Enem. Our study demonstrates the potential of the Progressive-Restricted method in adaptive testing to improve high-stakes exams administered in paper-and-pencil format by effectively controlling item exposure without compromising test precision. Future research should also validate these findings with real test takers.
Visualizing uncertainty to promote understanding of measurement error in score reports

Niek Frans, Benjamin Hummelen, Casper Albers, Muirne Paap

Measurement error is an inherent part of any test score. This uncertainty is generally communicated in ways that can be difficult to understand for practitioners in educational and clinical contexts. In this empirical study, we evaluate the impact of several communication formats on the interpretation of measurement accuracy and its influence on the decision-making process. We presented a sample of psychologists and remedial educationalists (N = 230) with score reports in five different formats (i.e., textual confidence intervals, error bar, violin plot, diamond plot, and quantile dot plot) and asked them to make probability assessments and referral decisions based on the information in the score report. Additionally, we asked participants to rank formats on understandability and describe their personal preference. We found that communication format was related to the accuracy of probability statements and that higher accuracy was moderately associated with better referral decisions. However, we could not find a direct relation between communication format and decision quality. Although many participants favored traditional confidence intervals and error bars due to their familiarity, responses revealed several misconceptions that make the suitability of these formats for communicating uncertainty questionable. Our results indicate that, even with minimal explanation, innovative visualization formats can successfully reduce errors in interpretation. With the addition of adequate training, such formats may eventually lead to higher quality decisions in education and clinical practice.
Investigating an invariant item-cluster ordering using nonparametric item response theory
Letty Koopman, Johan Braeken

Invariant item ordering (IIO) is a property of a test or questionnaire, in which the items have the same order in difficulty (or popularity) for all values of a unidimensional latent variable. An IIO allows for the arrangement of items from easy to difficult, better interpretation of test scores, and comparison of response patterns. Method manifest IIO and coefficient HT are useful tools for investigating an IIO in a test without placing other distributional restrictions on the items. For tests with clusters of approximately equally difficult items, the usefulness of method manifest IIO and coefficient HT is limited. For such a test, certain violations of IIO may be expected within clusters, whereas the clusters should be invariantly ordered. However, method manifest IIO will nominate items for removal even when violations are only present within clusters. Furthermore, the analysis provides no information on whether there is a meaningful ordering on the cluster level. In this presentation, we define an invariant item-cluster ordering, useful for tests in which items are presented in blocks for which the order in difficulty within blocks may vary across respondents, but in which the item-blocks are increasing in difficulty. We adapt method manifest IIO to allow for within-cluster violations of IIO and we propose a generalization of coefficient HT for assessing the accuracy of the cluster ordering. We apply the proposed analysis to data from the Norwegian version of the Test for Reception of Grammar. Suggestions for practice and further research are discussed.
Classroom Climate and Teacher Self-Efficacy in Culturally and Linguistically Diverse Elementary Classrooms: A State-Trait Analysis

Jacqueline Michelle Peterson

The present study is a longitudinal investigation which aims to assess the extent to which classroom climate and teacher self-efficacy manifest as stable traits (trait-like) versus fluctuating states (state-like) over time in culturally and linguistically diverse classrooms. Additionally, this study aims to explore whether class cultural and linguistic diversity, classroom climate, and/or teacher self-efficacy contribute to any observed changes in these constructs over time. In doing so, this study addresses current gaps in the literature in addressing the multidimensional constructs of classroom climate and teacher self-efficacy from a longitudinal perspective within culturally and linguistically diverse classrooms. Furthermore, in a departure from commonly used proportional measures of class cultural and linguistic diversity, the present study derives class cultural and linguistic diversity based on students averaged cognate linguistic distances, or lexical similarity to Norwegian, gleaned from inferred language background based on parents’ reported country of birth. This operationalization allows for greater precision in the measurement of class culture and linguistic diversity, as it positions class cultural and linguistic diversity in proximity to Norwegian culture and language, rather than treating students’ diverse backgrounds as a homogenous group. The present study uses longitudinal data pulled from the control group of a large randomized controlled trial intervention project, Two Teachers, in Southwestern Norway, as its data source. Specifically, the study investigates 2880 students, representing 95 different country backgrounds, nested within 150 control classes. Utilizing a latent-state trait perspective to longitudinal structural equational modeling, multitrait-multistate (MTMS) analyses within a multi-level framework will be used to investigate state-residual and trait variation in students’ reports of classroom climate and teachers reported self-efficacy across three consecutive years (2nd, 3rd, and 4th grades). After assessing variation, class cultural and linguistic diversity, classroom climate and teacher self-efficacy will be examined for their ability to explain variation in these constructs over time.
Large-scale assessments (LSA) rely on students' self-report questionaries to assess school environment effects. Scores from these scales are generated using unidimensional latent trait models, including variants of the partial credit model, and the graded response model. These models included students as a source of variance. As such, most scale quality is judged regarding students as a population, and reliability is judged regarding students' scores. However, the inference of interest of these scales are classroom attributes, rather than students. The disparity between scale interest, and the latent trait model used to generate its scores has at least two limitations: a) the response model ignores the cluster common shared variance, and does not generate the score of interest; b) the response model does not help to represent the measurement error in multilevel conditioning models, used to assess relationships between learning outcomes and classroom attributes. Previous work asserts school level factors can reverse their effect, when comparing aggregated means in contrast to latent means in the conditioning model (e.g., Pokoprek, 2015). We propose a mixed partial credit model, which allows us to generate cluster-level realizations, and students within school realizations, while considering students and classroom as relevant sources of variance of the response process. Using the proposed model, we can generate plausible values to include as part of the conditioning model in the outcome of interest to assess the relationship between the classroom attribute, and the learning outcome, while accounting for the measurement uncertainty of the contextual factor. We use data International Civic and Citizenship Education Study 2016, and Trends in International Mathematics and Science Study 2019, to illustrate the proposed application. We use Open classroom discussion, and Disorderly Behavior During Mathematics Lessons from each study respectively.
Effect of item responses and distractor choices on response certainties: an application of multilevel SEM.
Chia-Wen Chen, Jinxin Zhu

The response accuracy (i.e., giving correct answers), response certainty (i.e., confidence in answering items), and distractor choice (i.e., choosing a distractor due to misconceptions) in item responses are indicators for measuring cognitive ability, self-confidence, and misconception in mathematics tests. They have been investigated at between-person and between-item levels under the item response theory framework in literature. However, no studies examine a parametric model that simultaneously captures the relationship among the three indicators at the three levels, within-person-within-item, between-item, and between-person levels. This study proposed a cross-classified multilevel approach to address this research gap. Specifically, we applied a cross-classified multilevel model to investigate the relationships between response accuracy, response certainty, and distractor choice in a test with multiple-choice items at the within-person-within-item level, between-item level, and between-person level after controlling for the effects of background variables including response time, gender, prior achievement, and previewing the textbook before the answering the test paper (preview). The data were collected from 773 grade 7 students taking a mathematics test with 11 multiple-choice items. The results showed that at the within-person-within-item level, a high probability of giving correct answers would increase the response certainty. At the between-item level, item difficulty (response accuracy) and distractor efficiency (distractor choice) did not significantly relate to item certainty (response certainty). At the between-person level, student self-confidence (response certainty) positively predicted mathematics ability (response accuracy), whereas misconception (distractor choice) was negatively related to self-confidence (response certainty). The conclusions matched our expectations, and the implications were discussed.
Synthesizing Contextual Effects: A Comparison of the Two-Stage Meta-Analytic and One-Stage Multilevel Modeling Approaches in the presence of Individual Participant Data

Diego Campos, Mike W. Cheung, Ronny Scherer

Group composition effects are a central focus of social science research. Nevertheless, the generalizability of these effects across populations and settings remains unclear. Multilevel Models (MLM) and two-stage individual participant data (IPD) meta-analysis are two commonly used approaches to estimate and synthesize generalizable contextual effects. However, MLM with random slopes are computationally slow, imprecise, and frequently lead to convergence problems. Moreover, empirical evidence regarding the performance and accuracy of the two-stage IPD meta-analysis approach—a viable alternative to MLM—is still lacking. In this study, we compare and evaluate the statistical performance of the two-stage meta-analytic and one-stage multilevel modeling approaches for synthesizing contextual effects in social science research using individual participant data. We present a simulation study comparing the convergence rates, relative bias, RMSE, and observed coverage rates of the two approaches and a case study applying the two-stage meta-analytic approach to data from international large-scale assessments. The results provide empirical evidence on the performance of two-stage meta-analytic and one-stage multilevel modeling approaches for estimating and synthesizing contextual effects, which can guide researchers in selecting an appropriate methodology for their research.
CheckMate: using natural language processing to help teachers score short answers to open-ended items faster  
Eva de Schipper

For different reasons, teachers may want to administer tests that include open-ended format items. However, scoring open-ended format items is a time-consuming task that reduces the amount of time teachers can spend on other didactic tasks. As such, it is valuable to develop methods that can reduce the amount of time teachers spend on scoring these items without it being at the expense of scoring reliability. With this motivation, Cito Foundation developed a prototype named ‘CheckMate’. It uses techniques from the field of Natural Language Processing in order to compute similarity scores between different student answers and uses these to display similar student answers together, as well as make scoring suggestions for answers that have not yet been scored by the teacher. By making use of an open-source general language model, the system can be of use even if no earlier answers to the item have yet been scored. The current study represents a first step in validating the use of the developed prototype in educational practice with regard to scoring speed, quality of scoring, and user-friendliness through both qualitative and quantitative methods. In this presentation, I will demo CheckMate and elaborate on the results of the validation study, the plans for further development and research, and the wider applications of the techniques used in this prototype for the field of educational measurement.
In higher education, the use of essays as a formative assessment format is well suited to consolidating students' knowledge and stimulating reflective processes. Especially when combined with informative feedback, positive effects on student learning can be expected (e.g. average effect of $d = 0.99$ on learning success; Wisniewski et al., 2020). However, in large university courses with several hundred students, it is usually not possible to provide highly informative feedback using such test formats. One reason for this is the time-consuming scoring process. In addition to the reading time, trained coders are required for objective, reliable and valid scoring. Natural language processing (NLP) methods can be used to approximate the judgments of human coders to reduce the scoring effort. If the results are sufficiently accurate, they can be used as the basis for automatically generated feedback. The present study compares different NLP approaches for automated essay coding in terms of their agreement with human coders. The corpus consists of essays written by $N = 698$ students (max. 2 pages) in the context of a very large university course. First, a conventional content analysis was carried out by pre-trained coders. After the usual preprocessing steps, we compared three different NLP approaches (bag-of-words + random forest classifier, structural topic modelling + random forest classifier, transformer based BERT modelling) in terms of agreement with human scoring (precision, recall, F1 score). For most codes, the BERT modelling approach showed the best performance with mean F1 scores $> .8$. However, for some codes the structural topic modelling approach gives comparable results which are much easier to interpret and explain to students. The results are discussed with regard to validity issues in automated scoring engines.
AI-Literacy Network: Automatically analyzing stories written by primary school students in an international literacy network
Steinrücke Johannes, Amir Haeri Maryam, Roelofs Erik

The development of students’ writing proficiency is known to be supported strongly by individual teacher feedback based on formative assessment tools. Kennedy and Shiel (2021) recently reported on the validity and usability of a fine-grained rubric that addresses both lower (conventions, grammar) and higher-order levels (ideas, voice) of writing skills. However, using rubrics in large and diverse classrooms can increase the workload for teachers. The AI-literacy network, that is currently built in a European consortium of universities and primary school aims to foster free creative writing while making use of AI-technology that supports both teachers and students in supporting writing engagement and in providing feedback on the writing performance. In this sense, some of the teacher workloads is taken away. With the use of AI-techniques such as topic modeling and sentiment analysis, students’ written texts can be analyzed to provide formative feedback to the students. This may involve technical writing skills (spelling, syntax), but also analyses of higher-order text features (idea, organization, creative use of language). However, technology for such analyses, that could be based on automatically scored rubrics, are still lagging (Kumar & Boulanger, 2021). In this pilot study, written (and translated) texts from primary school children from three participating European countries are automatically scored using the formative rubric by Kennedy and Shiel (2011), applying supervised machine learning. In contrast to methods that only yield summative scores for feedback (Ke & Ng, 2019), using a validated rubric allows us to automatically provide formative feedback about both lower-order and higher-order levels of the writing product, including the idea and voice of the text. We will present preliminary results, challenges and recommendations.
Advancing DIF Analysis: A Three-Phased Approach for Identifying and Mitigating Bias in High-Stakes Assessments

David Budzynski, Ben Smith

Differential item functioning (DIF) analysis has become increasingly important in the field of assessment and evaluation as it helps to identify potential bias in assessments, which can have significant implications for individuals and groups. The aim of DIF analysis is to determine whether any demographic groups perform differently on each item in an assessment, after controlling for overall ability levels. However, identifying the presence of DIF alone does not help item authors understand how best to address bias in items. It is crucial to understand the root causes of DIF and to develop strategies to minimize its occurrence in future assessments. To attempt to address this, we developed a multi-phased approach for conducting DIF analysis. The approach consists of initial DIF analysis, focus groups with recent sitters and content experts aiming to identify explanations for DIF, and a series of regression analyses (based on item linguistics and metadata) aimed at establishing whether there are particular item features that tend to be associated with DIF. Through this innovative method, it is possible to not only identify which items exhibit DIF, but also highlight specific factors and item writing practices to consider when reviewing or writing items to mitigate DIF as much as possible. To summarise, this novel DIF analysis approach provides valuable insights into the factors that tend to contribute to DIF across suites of items and tests (rather than just individual items). It can thus offer effective guidance for test developers and policymakers on how to minimize bias in assessments. Our presentation of this approach will outline the methodology, its application in high stakes professional assessments in the United Kingdom, and seek input on future advancements and developments we have in mind from the audience.
Using Diagnostic Models to Evaluate Student Learning Hierarchies in a Large-Scale Assessment
W. Jake Thompson, Brooke Nash, Jeffrey C. Hoover

In recent years, the focus of large-scale assessments has begun to shift from summative evaluations of student achievement to more actionable and fine-grained information about student proficiency of skills within cognitive learning models. This change in emphasis not only has benefits for teachers, students, and parents, but also offers opportunities for researchers to better understand student learning and skill acquisition. To meet these challenges and opportunities, we need innovative psychometric models that we can use for reporting valid results and understanding student cognition. In this presentation, we describe a framework for using diagnostic classification models (DCMs; also called cognitive diagnostic models) within an operational assessment program to measure student proficiency and evaluate hierarchies in skill acquisition. The diagnostic framework for evaluating hierarchies includes multiple methods and applications of DCMs, each providing complementary information for evaluating the theoretically implied skill hierarchy and for identifying potential causes of violations to the hierarchy. We then illustrate the diagnostic framework in practice using data from a large-scale operational assessment, the Dynamic Learning Maps® (DLM®) Alternate Assessment System. Results for DLM assessments are calculated using DCMs to provide instructionally informative results within the proposed cognitive learning model structure. Thus, DCMs are used for both scoring and evaluating skill hierarchies within the DLM assessments. The findings demonstrate we can use DCMs for providing actionable assessment results and evaluating skill hierarchies to better understand student learning. Using the proposed DCM framework, we show that DLM skills largely conform to the hierarchy implied by the assessment design. Finally, we discuss how we can use findings to inform revisions to skill hierarchies and future avenues of research.
A New Framework for Universally-Designed Assessments
Meagan Karvonen, Cara Laitusis

The Standards for Educational and Psychological Testing (American Educational Research Association et al., 2018) describe universal design (UD) as a promising approach to promote fairness of educational assessments – for example, by designing assessments with a wide range of examinees in mind, or supporting flexibility during assessment administration. UD principles in education are adapted from UD concepts applied in the design of physical spaces. There are several applications of UD concepts to educational assessments (e.g., CAST, 2015; Lazarus et al., 2022), but each has its limitations. There are also inherent tensions between the flexibility needed to enact UD principles and the standardization valued in some forms of assessment. As the field reconsiders questions of equity and fairness in assessments (e.g., Randall, 2021) including culturally responsive assessment practices (Walker et al., 2023), we propose a new framework for universal design in assessment systems. This new framework is rooted in original UD principles from the field of architecture and incorporates contemporary perspectives on neurodiversity and culturally and linguistically relevant practices. The framework, with associated principles and guidelines, is intended to promote UD in coherent ways across assessment systems. The principles extend beyond assessment design and development to score interpretation and use and validity evaluation. This presentation will describe the theoretical foundation for the framework and how it addresses the gaps in current UD principles. We will describe how the principles and guidelines might apply in the context of large-scale assessment systems. The presentation will conclude with practical implications and suggested research priorities.
Developing the Basic Skills Assessment: a math test for a high dosage tutoring program
Kim Brandes, Milada Speet, Karmijn Steekelenburg, Thijs van Leuven, Gwendoline Moenandar, Samantha Bouwmeester

The aim of this multidisciplinary project is to develop a math test that can precisely measure the progress of children who participate in a high dosage math tutoring program (the Rekenfaculteit). The Rekenfaculteit offers 2 (children) to 1 (tutor) tutoring for four hours a week during the entire schoolyear to children aged between 8 and 10 years old in 11 primary schools in challenging areas in Rotterdam, the Netherlands. The Rekenfaculteit strives to improve children’s math skills. Math skills are routinely measured with national tests, like the CITO test. These national tests give global information about children’s math skills and allow for a comparison of the participating children with the national population by classifying children in the Netherlands in quantiles. However, these national tests are not suited to 1) provide detailed information of children’s mastering of specific math categories, 2) detect children’s insufficient math strategies and 3) reliably evaluate the progress of children in the lowest quantile. To tackle these issues, a multidisciplinary team was created (consisting of practitioners, researchers, experts in psychometrics and math didactics) to develop the Rekenfaculteit’s own math test: the Basic Skills Assessment. First, items with a relatively low difficulty level were created for all math categories and administered to a large number of students in Rotterdam. Second, the psychometric properties of the items were evaluated and high quality items were selected. We will present the results of our first year of data collection, our experiences with developing our own math test in a multidisciplinary team and specific plans for the second year of test development.
Implementing game-based assessment: the effect of context authenticity on measurement validity
Aranka Bijl, Sebastiaan de Klerk, Saskia Wools, Bernard Veldkamp

Administering standardized assessments through games has several advantages. Games allow us to collect rich, standardized data that can be used for stealth assessment, give the freedom to recreate many different environments, and provide a more enjoyable testing experience. But how do we make an assessment environment feel like a game, while ensuring that valid inferences about candidate ability can be made? Games have a number of defining characteristics. They are often characterized by having a clear goal that is given structure with rules, choices, and feedback. In addition, games often have a (fantasy) context that is different from everyday life. The context of the assessment environment thus likely affects whether candidates perceive it as a game. However, it is unclear whether the context would also affect the measurement validity. Therefore, this research aims to investigate whether the authenticity of the context within a game-based affects the measurement validity. This study focuses specifically on the assessment of problem solving because this is a skills that would benefit from a more generic and game-like assessment environment which can be administered across the many different contexts in which it is applied. The principles in Educational Design Research were used to develop an authentic assessment environment to assess problem solving skills of Dutch public safety officers in training. This was done in collaboration with subject matter experts and students, who provided input for each of the development iterations. Key elements that underly problem solving were then translated to two more game-like contexts that the target population is not familiar with, following the same principles of development. In the coming months, the authentic environment will be validated and performance (e.g., through in-game behaviors) between the contexts will be compared. Preliminary results will be discussed during the presentation.
Learning from mistakes - or not? Examining repeated mistakes in app-based learning.
Jarl K. Kristensen, Janne v. K. Torkildsen, Björn Andersson

In many educational apps, there is an element of trial and error, which means that learning from mistakes is key to facilitate the overall learning process. When learners keep repeating the same erroneous answers, it indicates a failure to learn from mistakes. In the current study, we examined repeated mistakes in a morphology-based app designed to enhance word knowledge in primary school children. 717 Norwegian second graders participated in a randomized controlled trial, of which 363 children worked with the morphology app during an eight-week intervention. We investigated 1) whether there is a change in the number of repeated mistakes over the eight weeks, and 2) how the number of repeated mistakes relates to pretest and posttest scores on a test of receptive word knowledge administered through the app. Repeated mistakes were extracted from process data log files. Each erroneous response chosen more than once during a task counted as a repeated mistake. Results from CFA analyses indicated that the number of repeated mistakes children make is relatively stable across sessions, and well represented as a unidimensional construct. Results from a SEM model where repeated mistakes mediated the effect of pretest scores on posttest scores indicated that children with lower pretest scores were more likely to repeat errors. Repeated mistakes in turn negatively related to posttest scores, indicating a negative impact on learning outcomes.
From Start to End: Item Nonresponse as a Function of The Sequential Item Position on the PISA 2018 Student Questionnaire
Kseniia Marcq, Johan Braeken

The Programme for International Student Assessment (PISA) student questionnaire is designed to be of low cognitive demand and easy to complete within the given time. However, the questionnaire’s length of about 300 items and the low stakes involved for individual students may result in a test burden that translates into a visibly increasing trend in item nonresponse towards the end of the questionnaire. Using a cross-classified mixed effects model and the PISA 2018 student questionnaire response data from 80 countries, this study examined the trend in item nonresponse as a function of sequential item position on the questionnaire and the extent to which this trend varies between individuals and across countries. On average, the probability of nonresponse was estimated at roughly 2% at the start of the questionnaire, gradually increasing until reaching 11% at its end, with a 12% spike in item nonresponse for two questionnaire scales. Considerable variation in nonresponse trend was noted between countries, with South American countries displaying a steeper increasing trend, whereas a flatter trend was observed in the Middle East, Asia, and the Nordics. Individual differences were equally prominent, with students who showed lower item nonresponse propensity at the start of the questionnaire exhibiting a steeper increase in nonresponse as they progressed, while those with a higher item nonresponse propensity at the beginning remained at a relatively constant level until the end. We discuss the implications of these findings for questionnaire design and researchers who use PISA student questionnaire data for secondary analysis.
Prevalence of Random Responders as a function of Scale Position and Questionnaire Length in the TIMSS 2015 eighth-grade Student Questionnaire

Saskia van Laar, Johan Braeken

This study examined the impact of two questionnaire characteristics, scale position and questionnaire length, on the prevalence of random responders in the TIMSS 2015 eighth-grade student questionnaire. While there was no support for an absolute effect of questionnaire length, we did find a positive effect for scale position, with an increase of 5% in random responding over the course of the questionnaire (in both the shorter and the longer version). However, scale character turned out to be an unexpected but more important determinant. Scales about students' confidence in mathematics or science showed an increase of 9% in random responding, which is double the impact of scale position. Potential mechanisms underlying the confidence case and general implications of the results for questionnaire design are discussed.
Test-taking behaviors after manipulation of test stakes
Michalis Michaelides, Despoina Chatzidimitriou, Myria Chatzimichail, Valentina Ierotheou

International, large-scale assessment programs, like TIMSS and PISA, are important for research and policy purposes with potentially high-stakes outcomes for the participating jurisdictions. However, they are considered low-stakes for the individual students who respond to the achievement tests. Individual scores are not reported, and so there are no consequences for the student examinees. Because of the low-stakes, previous research has identified decreased motivation and effort in these assessments (Eklöf, 2010). Test-taking effort correlates moderately to strongly with performance (Silm, Pedaste, & Täht, 2020). Applying a survey experiment paradigm, Zhao, Brown and Meissel (2020, 2022) compared test-taking behaviors across contexts. In a student sample in Shanghai they found no difference in self-reported test-taking effort, anxiety, and importance in the high versus low stakes (personal and country-stakes) conditions; but in a New Zealand sample, there were differences in those behaviors across conditions. The proposed study aims to replicate and extend this survey experiment in Cyprus. In a sample of 15-year-old students we investigate (a) whether there are differences in self-reported test-taking effort, anxiety, and importance beliefs when the test-stakes are manipulated as high-, low-, or no-stakes via survey vignettes; and (b) if test-taking behaviors relate to demographic and personality characteristics, and school type. Ethical permission to conduct the study has been obtained, and data are currently being collected from public and private Lyceums and Technical Schools in Cyprus. Analysis with structural equation modeling will include a comparison of self-reported effort, anxiety and importance beliefs across the survey experiment conditions, as well as associations with gender, school type, subject specialization, and conscientiousness scores. We hypothesize that test-taking effort, anxiety and importance beliefs will be higher in the high-stakes condition, and that test-taking self-reports will vary by school type, subject specialization and conscientiousness, particularly in the high-stakes condition.
Parallel optimal calibration of mixed-format items for achievement tests
Frank Miller, Ellinor Fackle-Fornius

When large achievement tests are conducted regularly, items need to be calibrated before being used as operational items in a test. Methods for assigning calibration items to examinees in an optimal way based on their ability have been developed. Most of these methods are intended for a situation where examinees arrive sequentially for being assigned to calibration items. However, several calibration tests are conducted in a way where all or many examinees conduct the test in parallel. In this talk, we develop an optimal calibration design for such parallel test setups which can be implemented in real calibration scenarios. In many real test situations, items are of mixed format and our optimal design method can handle that. We discuss first the optimal calibration designs for the 2-parameter logistic, the 3-parameter logistic, and the generalized partial credit model. Then, we consider the case of mixed-format tests consisting of items in all of these models. The method we propose can also take different expected solve times into consideration. We investigate the efficiency gain of the method. Our investigations show that the proposed method is able to increase calibration efficiency considerably. The described method has been used for calibration of items for Swedish national tests in Mathematics.
Large-scale achievement tests require the existence of item banks with items for use in future tests. The item parameters need to be estimated as efficiently as possible before being included in the bank. The precision of the item parameter estimates is depending on the estimated abilities of the examinees taking the items. Optimal design methods have been developed to allocate calibration items to examinees with the most suited ability.

In this presentation, we investigate the performance of an optimal ability-dependent allocation in the context of the Swedish Scholastic Aptitude Test (SweSAT) and quantify the gain from using the optimal allocation. We apply a method based on an optimal design algorithm that shows advantages in theory. The algorithm utilizes a so-called optimal restricted design using the D-optimality criterion. The computations involved in the optimal design algorithm use some approximations, and the optimal design needs to be evaluated in a real testing situation. We performed a simulation study to explore the optimal design strategy in a realistic setting. And we compared the results to a random design strategy, where items are instead randomly allocated to examinees. The evaluation measures used in the simulation study are developed carefully to mimic the measures used in optimal design theory.

We have quantified the gain from using the optimal allocation and, on average over all items, we see an improved precision of calibration. We are also able to identify for what kind of items the method works well.
Dealing with Local Item Dependence in item bank calibration and CAT administration: The item-enemy principle.
Johan Braeken, Nick Frans, Muirne Paap

A new pragmatic approach to deal with local item dependence (LID) in the context of computerized adaptive testing (CAT) is proposed and its effectiveness is shown.

The approach builds on the principle of "item enemies": If items part of a LID cluster are not jointly present, there is no way for the LID to express itself; From this perspective, items part of the same LID cluster can be seen as each other enemies, all potentially carrying unique information that would suddenly get artificially double-counted once the items are jointly present. This item enemy perspective suggests the following procedural steps. (i) Detect LID in the item bank that will be used to fuel the CAT; (ii) Apply an iterative item bank calibration scheme where for each calibration only one item of each detect LID cluster is included and the corresponding item enemies are excluded. This is repeated until all items in the LID clusters have been calibrated at least once and final item parameter estimates are obtained upon pooling the multiple calibrations. (iii) Run the CAT based upon the final pooled calibration of the item bank while applying the content constraints that once an item of a LID cluster has been administered, no other items in that cluster can be administered in the same test.
Anastasios Psychogiopoulos, Niels Smits, Andries van der Ark

Lengthy tests can lead to boredom and decreased motivation among respondents, especially those with a short attention span. To prevent biased assessment and incorrect or suboptimal treatment, computer adaptive testing (CAT) can reduce test duration without compromising measurement accuracy. However, traditional CAT models may not be suitable for many tests due to their demanding requirements. This study builds upon a flexible testing procedure, called FlexCAT, which accommodates various test formats and measurement levels. Specifically, we focus on the density estimation method for FlexCAT, using a latent class model (LCM) to estimate the joint density of item scores ($\pi$) and the density of the total score ($\pi^+$). We present the results of three simulation studies: (a) examining factors that affect the accuracy of $\pi$ and $\pi^+$ estimation using LCM, (b) comparing $\pi$ and $\pi^+$ estimation between LCM and models commonly used in CAT, and (c) presenting an example of a complete FlexCAT simulation.
Semi-adaptive testing with block design – comparing design-determined and estimated item difficulties
Daniel Bergh, Tine Nielsen

In recent years significant developments have been achieved within the area of Computer Adaptive Testing (CAT). In CATs, item difficulties are central, as they are matched with individual’s ability during the test situation, ideally resulting in precise measurement (i.e. SEM less than 0.3 for all persons). Not all CATs are fully adaptive in the sense described, but instead employ a semi-adaptive approach, where sequences or groups of tasks are distributed to persons at a specific ability “interval”. The purpose of this study is to analyze data collected using a semi-adaptive approach with a block design. Specifically, the study is aimed at discovering how the designated difficulties from such an approach are related to the estimated item difficulties that would be used in a full CAT design. Data was collected among persons assessed for conscription by The Swedish Defence Conscription and Assessment Agency (SDCAA) using the designated test of general cognitive ability (i.e. intelligence) (N = 1469). In this study, two subscales of the cognitive ability test were analyzed by the Rasch Model, and item difficulties were estimated. Item difficulties were then grouped into three levels (easy, medium, and hard) in order to facilitate comparisons with the three difficulty levels in the semi-adaptive test design used by the SDCAA. The analysis revealed discrepancies between the estimated item difficulties and the beforehand established difficulty grouping used in the semi-adaptive design. The adaptive test design will be presented, and the consequences of these discrepancies will be illustrated and discussed. The implications of the findings for educational contexts beyond that of military education will also be touched upon.
The Link between Gender Gaps in School Enrollment and School Achievement

Isa Steinmann, Leslie Rutkowski

Usually, gender gaps in school enrollment and achievement are investigated as separate facets of gender inequality in education. This study suggests that they might be linked by design. If boys and girls differ in their school enrollment rates, the student population will be gender biased. Since out-of-school adolescents tend to be less advantaged than in-school ones, school-based large-scale assessments effectively compare a rather fully represented group of one gender with a less represented and more advantaged group of the other one. This should shift student gender achievement gaps so that they favor the latter group. In country-level regression models using data from all PISA (Programme for International Student Assessment) cycles and PISA for Development, we indeed found evidence of a small, negative association between gender gaps in secondary school enrollment and gender gaps in student achievement. The school enrollment data stem from UNESCO (United Nations Educational, Scientific and Cultural Organization) indexes. This finding is robust across different achievement domains and specifications. This study illustrates implications of the fact that the school-based assessment PISA only covers students, not out-of-school youth. The implications for the interpretation of PISA findings will be discussed especially considering the finding of international variations in achievement gender gaps.
Behind the scenes: The Transition from Paper-Based to Technology-Based International Large-Scale Assessments
Heiko Sibberns

International Large-Scale Assessments (ILSA) conducted by the IEA have been in transition from paper-based to technology-based assessment (TBA) for some time. This transition entails massive challenges for both the international study centers and the national study centers, which are hardly visible to the end users of the data at first glance. This applies to changes in the work processes that are carried out when implementing a study, in the time schedules, changes in the required resources and new technical infrastructure, shifts in work processes from national to international study management, and new quality control procedures - to name just the most important. This goes hand in hand with changed knowledge and skills for the required employees and, finally, shifts in the allocation of financial resources. These changes take place inconspicuously in the background but can have consequences for the data collected with the help of technology-based assessment.

In the presentation, the work process from the planning to the implementation of the assessment under TBA conditions should be presented. The difference to the traditional paper-based assessment should be shown. It also explains what consequences may arise for the collected data and what end users should pay attention to. The focus will be on the significantly higher documentation effort, which places new demands on the data provider and means additional effort for the data user. No specific technology-based assessment system is considered. However, the presentation will be limited to systems that have been developed or are in use as part of IEA studies.
Using TIMSS to evaluate the outcomes of an alternative mathematics curriculum
David Greger, Eva Potužníková, Patrícia Martinková

Large-scale assessments provide representative data on student learning outcomes to inform decision making and research. Despite an expansion of national assessment programmes (OECD, 2013), international large-scale assessments (ILSAs) remain an important data source not only owing to high-quality achievement tests but also due to rich contextual data. This study uses TIMSS 2015 data from the Czech Republic (4,790 fourth grade students nested in 243 classes) to evaluate the outcomes of an alternative mathematics curriculum that has been recently implemented in some primary classes across the country. This alternative curriculum, known as the Hejný’s method (Říčan, Chytrý, & Medová, 2022), stems from constructivism and uses specific learning tasks that purportedly support students’ reasoning skills (Artigue et al., 2020). We first estimate the effect of the Hejný’s method on the overall mathematics achievement with two-level regression models that control for student SES (Sirin, 2005), interest in mathematics (Gilbert, 2016), and instructional and teacher quality (Blömeke, Olsen, & Suhl, 2016). We then compare student achievement on subscales constructed by IRT scaling based on an expert classification of released items into three groups (i.e., items assessing skills emphasized by the Hejný’s method, items assessing skills emphasized in the standard mathematics curriculum, and items assessing skills covered evenly by both curricula). While the overall achievement of students taught according to the Hejný’s method is slightly, though significantly, higher than that of their peers, there are notable differences at the subscale level, indicating higher effectiveness of the alternative curriculum in developing specific skills. The study shows the potential of using ILSAs data to answer national education policy questions, but also points to the limitations, e.g., due to small number of Hejný’s method classes in the TIMSS sample.
Dimensionality of TIMSS science assessment: are subscale scores sufficiently distinct?
Yuan-Ling Liaw, Rolf Strietholt

The study aims to investigate the dimensionality of TIMSS science assessment and the distinctness of the subscale scores with a focus on the newly constructed environmental awareness scale. We specify a unidimensional IRT model with theoretically and empirically derived multidimensional IRT models. We evaluate the reliability, validity, and distinctness of the subscores when minor alternate dimensions are explicitly modeled. The increasing urgency of environmental crisis, and growing student concerns about the climate action, motivated international large-scale assessment research to focus on students’ knowledge surrounding the environmental and sustainable development. However, the so-called “environmental awareness scale” in TIMSS 2019 was established post hoc after test administration. Many science items addressing environmental issues were identified among the Earth and life science items at the fourth grade and the Earth science and biology items at the eighth grade. TIMSS therefore applied an alternative way of specifying content domains to construct two new scales, an environmental science scale and a nonenvironmental science scale without the environmental items, while keeping the original grade 4 physical science and grade 8 chemistry and physics scales unchanged. Recognizing this deviation from the standard approach, we examine the dimensionality of the assessment and determine whether the subscale scores provide added value beyond the general science score. In this study, three competing models are specified: (a) a unidimensional IRT model; (b) two theoretical bifactor models, where the dimensionality aligned with the assessment framework corresponding to the original content domains and the other with the alternative framework specifying environmental items, respectively; (c) an empirically derived model, with items found to load on alternate dimensions during the exploratory factor analysis. We determine if the resulting subscale scores based on the competing models are of added value. The practical and theoretical implications of the findings as well as the limitations will be discussed.
Posters

- The holistic and fragmented assessment of reading fluency in early primary education: A scoping review
  Max van der Velde, Bo Molenaar, Bernard Veldkamp, Remco Feskens, Jos Keuning

- Using Reaction Time Data to enhance PISA questionnaire indices
  Morten Rasmus Puck

- A systematic literature review of the Teaching and Learning International Survey (TALIS) peer-reviewed journal articles
  Jelena Veletić, Ana María Mejía-Rodriguez, Rolf Vegar Olsen, Sigrid Blömeke

- The Roles of Cognitive Abilities and Personality Traits in Inconsistent Responding to a Mixed-worded Scale
  Jianan Chen, Johan Braeken

- Application of the diffusion drift model to a dot comparison task for understanding quantity processing
  Jil Brambrink, Olga Kunina-Habenicht, Jörg-Tobias Kuhn

- Accuracy of short variable-length CATs
  Angela Verschoor

- Mathematics and Academic literacy cross-domain analysis in high-stakes assessment for augmented practices in diagnostic interpretation
  Tatiana Sango, Sanet Steyn

- Intrinsic Reading Motivation in PIRLS: A Rasch-based validity study
  Morten Pettersson

- Identifying PISA test disengagement using timing data and decoy items
  Tony Tan, Bryan Maddox
• Writing argumentative skills measurement in Peruvian students
  Luz Huanca

• Measurement invariance across educational systems in the First and Second
  International Science studies
  Yuriko Sosa Paredes, Björn Andersson, Sigrid Blömeke

• Wiener Diffusion Model - Estimation Problems and Dependencies on Initial
  Parameters
  Michael Radloff, Rainer W. Alexandrowicz

• Challenges in using Content Coverage in TIMSS as a mediator of changes in
  achievement scores.
  Monica Rosén

• Models with plausible achievement score values, missing background data, and FIML
  estimation: A software comparison
  Anne Traynor