Abstract book of the EduθMetric 2014 International Seminar on Educational Research and Measurement

Presentations in Session E1:

Artur Pokropek (artur.pokropek@gmail.com)
Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw
Educational Value Added Unit, Educational Research Institute, Warsaw

Application of Mixture Item Response Model for twin data with unknown zygosity
Keywords: heritability, IRT, latent class, mixture modeling, twin studies

Statistical modeling using twin data is increasingly popular in health sciences and psychology where they provide evidence on genetic and non-genetic causes of diseases or individual psychological traits. In social sciences twin studies aim at addressing the "nature or nurture". In educational research twins studies might be used to estimate effects of family background on educational outcomes. Most criticism referring to twin studies aims small sample size and lack of representativeness to the population of interest. This presentation shows how methodologies developed for twin data can be applied to information other that twin registers. It is shown how compulsory exam data along with IRT latent class modeling, can be, under some circumstances, used to generate reliable quasi-genetically informed designs, whereby estimates of heritability, common environment and specific environment might be obtained. Proposed model consists of three parts: (1) measurement part, (2) structural part and (3) latent class part. The measurement part is defined by two-parameter logistic Item Response Theory model. The core of structural part is the ACE. Because the zygosity is unknown in our data, a latent class part of the model is employed to estimate it. Proposed model was validated on data with known zygosity and proved to be highly reliable. Results on Polish examination data show high heritability estimates indicating that between 60-90% of variation in achievement may be attributed to genetic factors.

Maciej Koniewski (m.koniewski@ibe.edu.pl), Przemysław Majkut
Educational Value Added Unit, Educational Research Institute, Warsaw
Institute of Sociology, Jagiellonian University, Krakow
Paulina Skórska
Student Performance Analysis Unit, Educational Research Institute, Warsaw
Institute of Sociology, Jagiellonian University, Krakow

Detecting and profiling examinees with aberrant response patterns on lower-secondary school exit exam in Poland

Keywords: persons fit indices, aberrant response patterns, students characteristics, high-stake tests

Statistical procedures for detection of aberrant response pattern have been extensively studied (Cizek, 1999). Many person-level indices have been developed to detected examinees with spuriously high or low score (Karabatsos, 2003). However, little attention has been given to the characteristics of examinees with aberrant response pattern. Polish external exam data combined with large scale survey on lower secondary school students provides the unique opportunity to examine this issue. The research goal was to compute persons fit statistics and compare their values across , math, science and history tests. These four tests comprise high-stake exam administered on lower secondary school graduates (16 y.o.). It was assumed that this approach allows to indicate whether aberrations in response patterns are constant across test for particular examinee and therefore, aberrant response pattern is a person attribute rather that caused by a test-situation. Secondly, profiles of persons with aberrant persons patterns were prepared, based on demographic and contextual information. The person fit index lz (Dragsow, Levine and Williams, 1985) was computed for population data of 400K examinees, who undertook high-stake test after lower secondary school in 2012. The results has been linked to the demographic and contextual data gathered via survey on nationwide representative sample (ca. N=4900) of lower secondary school students. The multilevel logistic regression analysis was run to checked relationship between demographic and contextual students characteristics with person fit index.

Karolina Świst (k.swist@ibe.edu.pl)
Student Performance Analysis Unit, Educational Research Institute, Warsaw

The four-parameter logistic model – a useful tool in educational research or a conceptual humbug?

Keywords: creative responses, four-parameter logistic model, HT statistics, person-fit, teaching to the test

The four-parameter logistic model (4PLM) assumed that even the students who possess high level of ability, can make mistakes - not caused by their lack of knowledge, but for example carelessness. In this model, the upper asymptote of logistic curve is also estimated. The research on psychometric properties of 4 PLM was somehow hampered, as the model was conceptually and computationally complicated. What is more, the interpretation of upper asymptote can vary – as the wrong answer of high-achiever can be caused also by his creative response, not present in the scoring key. The creativity or unusual responses of high-ability students might therefore be punished. In this analysis, I am using 4PLM model in order to analyze psychometric properties of standardized large-scale exams on Polish language at the end of lower secondary school (from 2002 to 2014). Standardized exams are often criticized, as students are said to be 'taught to the test' and answer as close as possible to the scoring key. I am using mirt package to estimate 4PLM results and HT statistics from PerFit package in order to say whether the problem of 'wrongly punished' by standardized tests results high-achievers in fact exists. I am showing: a) how many items

exhibit upper asymptote b) how does the choice of distractors look like in flagged questions c) whether the choice of the answer can be interpreted as creative response and d) are the results of 4PL model consistent with the results obtained by HT statistics. Answering these questions is going to allow me to evaluate the usefulness of 4PLM in educational research.

Tomasz Żółtak (t.zoltak@ibe.edu.pl), Grzegorz Golonka
Educational Value Added Unit, Educational Research Institute, Warsaw

Does guessing matter? Differences between ability estimates from 2PL and 3PL IRT models in case of guessing

Keywords: ability estimate, guessing, IRT, 2PL, 3PL

In contemporary tests, used to measure cognitive abilities and knowledge, items in multiple-choice format are used very frequently. These can be very easy and fast scored and scoring is not prone to the problem of raters subjectivity. Nevertheless multiple-choice items are often criticized because of its vulnerability to guessing. It should be noted however, that guessing is a much more serious problem in criterion-referenced testing than in norm-referenced testing. In this paper, using simulation study, we examine impact of guessing on properties of ability estimates obtained from two IRT models that are widely used for binary-scored items: two parameter logistic model and three parameter logistic model. The first does not take into account possibility of occurrence of guessing, while the second explicitly models guessing. The question we ask is to what extent choosing one of these models affects inference about test-takers. To the best of our knowledge this problem has not been studied extensively before. To answer these questions we carried out simulations in which test results generated according to 3PL model were then calibrated using both 2PL and 3PL models and EAP ability estimates were obtained from these models. We examined correlation between obtained ability estimates and true ability used to generate test results and linearity of these relationships. Additionally we analyzed differences between estimates of standard errors of ability estimates obtained from 2PL and 3PL models. We studied these relationships in different conditions according to sample size, number of items in test and intensity of guessing. Rather counterintuitively we find very little difference in point estimates of ability obtained from 2PL and 3PL model. For relatively small sample sizes estimates from 2PL models turn out to be even marginally more strongly correlated with true values of ability than estimates from 3PL models. With large sample size this relationship reverses in favor for 3PL model, but differences are still very small. Nevertheless it must be noted that difficulty and discrimination parameters are severely downward biased if 2PL model is used to calibrate data generated by process involving guessing. Also estimated standard errors of ability estimates differ considerably between these models.

Presentations in Session E2:

Marek Muszyński (m.muszynski@ibe.edu.pl)
Foregin Language Section, Educational Research Institute, Warsaw
Institute of Psychology, Jagiellonian University, Krakow
Maciej Jakubowski
Faculty of Economic Sciences, Warsaw University, Warsaw

Cognitive strategies and reading outcomes: analysis of the PISA 2009 results for Poland

Reading skills are a key to learning and an indispensable „first step" to acquire any knowledge, thus reading development is often in the center of educational research. PISA results show that some unsettling gaps in reading skills exist not only across countries, but also within them. For example, between boys and girls or between students from different socio-economic background. One of the ways to narrow this gap might be to teach students how to use effective reading strategies. Existing research shows that using right strategies is associated with greater reading enjoyment and better reading performance (OECD, 2010). However, little is known which strategies are more effective for different groups of students. Moreover, there are so many proposed strategies that more empirical evidence is needed to avoid confusion and enable informed choice of the most effective ones (Dunlosky et al., 2013). To get new insights into the effectiveness of various reading strategies the PISA 2009 data for Poland was analysed. The effort was made to study influence of different strategies on reading outcomes and influence of those strategies depending on gender, socio-economic background or reading achievement level. To address these research questions techniques of quantile regression and multilevel models are applied. The results suggest that some strategies (summarising, understanding and remembering) are more effective than the others and some can be even counterproductive (memorisation). Those results are in line with some previous research (Chiu et al., 2007; Lau & Chan, 2003). Moreover, reading strategies were identified as a useful tool to narrow the gap between different groups of students or schools, while evidence demonstrates that the same strategy might work very differently depending on which group of student it is applied to.

Robert Mazelanik (robert.mazelanik@uj.edu.pl)
Institute of Sociology, Jagiellonian University, Krakow
Maciej Jakubowski
Faculty of Economic Sciences, Warsaw University, Warsaw

Reading Gender Gap and the lack of a father at home

In this study, we check whether the lack of a father at home is associated with a performance difference in reading between boys and girls. Our analysis explores data from the PISA 2009 study. A usual finding with these data is that girls outperform boys in reading in all countries, while the size of the difference varies across countries and social groups (Grey et al., 2004; Legewie and DiPrete, 2012; Machin and McNally, 2005; Jakubowski and Borgonovi, 2012). Similar findings are confirmed by other studies or assessments, for example, similar gap is present on the Polish national exams. According to PISA Poland has also one of the largest gender achievement gap across participating countries, mostly because of unsatisfactory performance of many boys. Most explanations of gender differences in academic achievement consider factors like psycho-biological differences, gender stereotypes, impact of parents or teachers as main the chief contributors to the gap. Several studies analyzed how the absence of a father is related to child's behavior,

well-being and educational outcomes (Albertini and Dronkers, 2009; Bernardi and Radl, 2014; Biblarz and Raferty, 1999; Jonsson and Gähler, 1997; Martin, 2012). However, we are not aware of a study that examines how the lack of a father relates to reading achievement gap and how this relationship depends on the student gender. The study aims to analyze (1) whether the absence of a father is related to a lower reading performance even after controlling for the socio-economic status, and (2), whether this effect is stronger for boys. In this regard we use data from the PISA 2009 study. We apply multilevel regression models to analyze the relationship between student performance in reading by gender and by family situation, while at the same modelling student- and school-level relationships and taking into account clustered sample design. The results suggest that students declaring themselves as living without fathers achieve on average lower results in reading, although the difference is rather small. In Poland, however, this effect is much stronger resulting in a reading score lower by more than 25 points among students declaring absence of a father. The most interesting result, however, is that father's absence has a stronger negative effect for boys. Across all countries boys declaring living in families without a father achieve scores that are lower than girls. In Poland, the difference is even larger. We hypothesize about possible sources of this gap and its much larger size in Poland.

Paulina Skórska (p.skorska@ibe.edu.pl)
Student Performance Analysis Unit, Educational Research Institute, Warsaw
Institute of Sociology, Jagiellonian University, Krakow
Karolina Świst
Student Performance Analysis Unit, Educational Research Institute, Warsaw
Aleksandra Jasińska-Maciążek
Educational Value Added Unit, Educational Research Institute, Warsaw
Maciej Koniewski
Educational Value Added Unit, Educational Research Institute, Warsaw
Institute of Sociology, Jagiellonian University, Krakow

Gender differences in pseudo(guessing) and item omission

Key words: gender gap, high-stakes test results, Item Response Theory, omitted items, pseudo-guessing

The gender gap in the results of large-scale standardized tests is of interest to researchers and stakeholders of the education system for several decades. Studies examining gender gap in tests results usually refer to the US population (e.g. SAT). Cross countries comparisons usually refer to the low-stakes tests (e.g. PISA). Such tests may result in low validity, as test-takers have low motivation to solve the test the best they can. Comprehensive explanations of boys-girls differences in performance on the high-stakes exams are required. In this paper tests and items characteristics are considered as possible source of boys-girls differences in performance on the exam administered at the end of the lower-secondary school. Tests on the 2012-2014 time span were analyzed. Two test taking strategies are of our main interest: (pseudo)guessing and omitting items. These strategies can differ among boys and girls, which can be explained within the same theoretical framework. In this study IRT modeling was used, especially the 3PLM for dichotomous items and GRM for polytomous items. The results show gender differences among boys and girls in a) the scale and effectiveness of (pseudo)guessing strategy, b) its interaction with item difficulty and c) tendency to omit items. The results may contribute to better understanding of gender gap in high-stakes tests results.

Krystian Barzykowski (krystian.barzykowsk@uj.edu.pl), Joanna Grzymała-Moszczyńska, Marianna Król
Institute of Psychology, Jagiellonian University, Krakow

Investigating gender differential item functioning in lower high shool exam. Example derived from 3 voivodeships: Lesser Poland, Subcarpathian and Lublin

Key words: DIF, educational measurement, gender differences, test design, IRT

Since 1999, when the educational reform establishing lower high schools called “gymnasia” took place in Poland, large-scale assessment plays a significant role in the Polish educational system. In Malopolska, Podkarpackie and Lubelskie voivodeships, every year over 60 000 pupils take up a final exam, results of which determine if they will be accepted to the high school they chose to attend. In this context one of the most crucial challenges is constant monitoring and detection of possible differential item functioning (DIF). According to Zumbo (1999), "DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure."Gender, generating the most basic division into groups, finds a constant interest of educational measurement researchers (Le, 2006; Liu, Wilson and Paek, 2008; Ryan and Chiu, 2001; Sireci and Mullane, 1994). However, the biggest number of research in the aspect of gender is conducted in the context of language and mathematics exams. The aim of this study was to test the potential presence of DIF in the 2014 science exam that integrates both verbal and computational skills. A multidimensional Rasch model in the TAM package in the R software was used for item calibration and ability estimation on the basis of 4 domains derived from PISA math dimensions: Space and Shape, Change and Relationships, Quantity, and Uncertainty (Shiel et al., 2007). Results showed that the effect sizes of performances for DIF between gender groups vary between small, medium and large, from 0.03 up to 0.77, depending on the dimension. The results are discussed in the context of practical implications for education testing and test design.

Presentations in Session L1:

Dennis Tamesbergera (tamesberger.d@akooe.at)
Department for Economic, Welfare and Social policy, Chamber of Labour, Linz, Johannes Kepler University, Linz

A multifactorial explanation of youth unemployment and the special case of Austria
Keywords: Youth unemployment, Austria, Cross-country analysis, active labour market policy, unions, vocational training, labour market regulation

One of the biggest challenges society currently is facing is the dramatically high level of youth unemployment. Political solutions and strategies can be especially found in those countries that have been able to keep youth unemployment low despite financial and economic crises. Austria is such a case. On the basis of EU member state data, this article gives a multifactorial explanation of youth unemployment and answers the question which problems in comparing youth unemployment exist, which factors influence youth unemployment and can these factors explain the relatively low youth unemployment in Austria? Further an overview of Austria's "youth safety net" is presented. This article points out both the youth unemployment rate's cross-country comparison limitations and that more indicators are necessary. Finally, this paper argues for economic policies to stimulate demand, which have to be based on a political and social commitment for full employment.

Zuzanna Drożdżak (zuzanna.drozdzak@uj.edu.pl)
Center for Evaluation and Analysis of Public Policies, Jagiellonian University, Krakow

Measuring socioeconomic position using categorical principal components analysis - the role of income, education and occupation for socioeconomic position in Poland
Keywords: ESS, optimal scaling, SES, social stratification

Many authors interested in the relationship between social inequalities and educational performance look for guidelines as to the best empirical measure of socioeconomic position (SES). I briefly review different conceptual propositions of how to measure family's position in a societal structure. Following that I formulate a list of requirement that a measure of SES needs to fulfill in order to serve well practical measurement purposes. The empirical part shows an attempt to construct such measure. To identify minimal data requirements for a reliable measure of socioeconomic position Polish edition of a cross-sectional European Social Survey, Wave 6 (2012 year) was used to develop three competitive measurement models of SES: These models were: "simplistic" (aggregating equivalized income and education), "simple" (equivalized income, education and occupation) and "rich" (equivalized income, education and occupation, supervision, influence on organization of own work and company's operations). Categorical Principal Component Analysis (CatPCA) allowed to stratify respondents based on these characteristics. Equivalized income and education build a one-dimensional model of socioeconomic position. Adding information about occupation and other job characteristics result in generating more dimensions, yet does not change the stratification of individuals on the first dimension of the the models. Spearman rank correlation between stratification obtained from the three measurement models was ranging from 0.905 to 0.972. Combined equivalized income and education is simple yet accurate measure of socioeconomic position in Poland.

Philipp Gerhartinger (gerhartinger.p@akooe.at)
Upper Austrian Chamber of Labour, Linz, Johannes Kepler University, Linz

Maternal working conditions and their influence on socio-emotional child development
Keywords: attachment, child development, maternal employment, stress process, working conditions

Research on the relationship of maternal employment and socio-emotional child development has not yet yielded homogenous results. There is little consensus in the scientific community on whether a positive or a negative relationship exists between the two factors. One reason for that can be that major variables have been ignored in the underlying theoretical models used. One such variable is the 'quality of work'. Within a stress theoretical model, however, the working conditions a mother is faced with are to be seen as a relevant factor. The objective of the presented study was to shed light on the indirect and direct correlations of maternal employment as well as maternal working conditions in the first three years of a child's life on the one hand and socioemotional child development in the last year of kindergarten on the other hand; all embedded in a stress and attachment theoretical model. In order to answer the research question a path analysis has been used based on data obtained from the Upper Austrian kindergarten attendants as well as their mothers. Child development was measured by a standardised development questionnaire which was filled out by the pedagogic kindergarten staff. The rest of the data was collected via a retrospective questionnaire filled out by the mothers of the children in the sample. The study has shown that maternal working conditions are a relevant factor in the relationship of maternal employment and socioemotional child development. While the working hours and/or point of (re)entrance into the labour market neither directly nor indirectly correlate with the development, the working conditions do negatively affect the stress process and therefore the level of the mother's stress. This in turn positively affects mother-child interactions, mother-child attachment and subsequently child development. Good working conditions therefore indirectly and positively correlate with the socio-emotional child development. In future studies interested in the relationship of maternal employment and child development, maternal working conditions should merit further and more in-depth consideration. Furthermore, there is strong evidence that maternal employment per se is not detrimental for child development, as it is, at least in Austria, commonly assumed.

Christina Koblbauer (christina44@gmx.at)
Department for Economic, Welfare and Social policy, Chamber of Labour, Linz, Johannes Kepler University, Linz

The institutional childcare situation of small municipalities in Upper Austria
Keywords: Institutional Childcare, Small Municipality, Upper Austria, Family Policy

On the one hand, the Austrian Familiy Report (Bundesministerin für Wirtschaft, Familie und Jugend 2010) shows that labour participation rates for mothers increase. On the other hand, reports (e.g. Bundesministerin für Frauen und öffentlichen Dienst 2010) show that opening hours of childcare provisions decline in rural areas and that it is difficult for parents to reconcile family and working life. Nevertheless other reports (e.g. Kammer für Arbeiter und Angestellte für Oberösterreich 2012) exhibit that some of those small municipalities in rural areas offer more extensive childcare provisions than others but do not pay attention to the reasons for this variance. The "Kinderbetreuungsatlas" (Kammer für Arbeiter und Angestellte für Oberösterreich 2012), which lists all childcare provisions of municipalities in Upper Austria expose differences between small communes in Upper Austria but does not consider influencing factors that cause variations. The objective of this proposal is to examine the extent of disparities in childcare provisions and use of these services between small municipalities and which factors affect the provision and

usage of childcare services. In a secondary data analyses the impact of certain characteristics of communes on the provision and usage of institutional childcare services has been tested. Therefore all municipalities in Upper Austria with less than 1000 inhabitants have been included in the study. A significant relationship between institutional childcare provisions and the political situation in communes, educational level of women and structure of family have been found. Among other things it is necessary that political parties start to rethink their attitude concerning institutional childcare in order to adapt to new societal demands.

Presentations in Session E3:

Przemysław Majkut (p.majkut@ibe.edu.pl)
Educational Value Added Unit, Educational Research Institute, Warsaw
Institute of Sociology, Jagiellonian University, Krakow
Gabriela Czarnek, Piotr Dragon
Institute of Psychology, Jagiellonian University, Krakow

Comparison of several approaches to analysis and shortening of psychological scales

Item-Response Theory (IRT) is getting more popular in the area of psychological scales development. We used IRT to shorten existing self-reported psychological scale that is often used in the domain of social cognition. The Need for Cognitive Closure (NFC) Scale (Webster and Kruglanski, 1994; Kossowska, 2003) is one of the most popular measures of cognitive closed-mindedness exhibiting promising psychometric properties. It is a 32-item scale where participants rate how they agree with presented statements on a 6-point Likert scale. In several studies researchers used selected items from the NFC scale without explicitly stating the reasons for that (e.g., Keller, 2005; Kemmelmeier, 2010; Lynch, Neteme, Spiller i Zammit, 2010). Recently however, the Polish version of a scale was shortened using Confirmatory Factor Analsysis (Kossowska, Trejtowicz, and Hanusz, 2012). In our research we aimed at providing the list of the most optimal items for shortened version of the scale using IRT modeling. Aggregated data (N=1754, Cronbach's alpha = .86) from 13 studies were analyzed using Graded Response Model (Sameijna, 1969). The criterion for choosing items to the final pool was their information function. Then, construct equivalence of shortened and full-length scale were checked. In our presentation we will discuss the usefulness of IRT as a method for shortening self-reported measures and compare it with traditional ways of scales reduction (i.e., Confirmatory Factor Analysis).

Agnieszka Strojny (agnieszka.strojny@uj.edu.pl)
Institute of Psychology, Jagiellonian University, Krakow
Kinga Karteczka-Świętek, Beata Nosek
Institute of Pedagogy, Jagiellonian University, Krakow

Difficulties and inaccuracies in translation and cross-cultural adaptation of items on the example of the international PISA study

Keywords: cross-cultural adaptation, item translation, test equivalence, PISA study

The authors consider the issue of difficulties and inaccuracies in translation and cross-cultural adaptation of the items. Justification of the need of international assessment programs and a brief description of the PISA study is presented. Based on the literature review the theoretical, social and translation equivalence issues are discussed together with an indication of what can be done towards achieving equivalence when creating international tests. The aim of the research was verification for correctness of translation and cultural adaptation of the 2009 PISA questionnaire. Analyzes of the distributions of student performance in selected countries for different items of the questionnaire were conducted. As a result, the authors discuss specific examples of the identified errors. The possible consequences for the accuracy of analyzes based on data collected using improperly prepared items have also been indicated.

Maciej Taraday (maciej.taraday@uj.edu.pl)
Cognitive Science Department, Jagiellonian University, Krakow
Anna Maria Wieczorek
Nencki Institute of Experimental Biology, Polish Academy of Sciences, Warsaw

Convergent validity of Raven's Progressive Matrices Test within Classical Test Theory and Item Response Theory

Keywords: classical test theory, convergent validity, fluid intelligence, item response theory, Raven progressive matrices

Precise measurement of fluid intelligence (Gf, Spearman, 1927) is crucial for the purpose of diagnosis as well as in case of scientific research on the nature of intelligence. Gf is the most basic component of intelligence. It reflects the ability to spot complex and abstract relations between symbols, as well as the ability to manipulate them. Due to Gf people are able to cope with similar problems which appear in different contexts and resolve them by using analogy. In this paper we discuss the most widely recognisable way of Gf assessment by using sum of correct answers in advanced version of Raven's Progressive Matrices Test (RPM, Raven, 1983) as an estimator of Gf. This method seems to be biased due to the fact that it does not take into account difficulty of the test items. Authors reanalysed data from 1724 subjects performing RPM (Chuderski, 2013; Chuderski, Taraday, Nęcka and Smoleń, 2012) and Figural Analogies Test (FAT, Orzechowski and Chuderski, 2007). FAT was chosen as a criterion task because it measures the most important aspect of human intelligence - ability to resolve problems by using analogy. Convergent validity of RPM and Figural Analogies Test (Orzechowski and Chuderski, 2007) was estimated by using raw sum of scores, factor scores from classical test theory (Gulliksen, 1950, in: Hambleton , 1993) and factor scores taken from 1 to 3 parameter item response theory models (Hambleton, 1991). Authors observed the highest correlation between criterion task and factor scores based on the item response theory in comparison to scores based on classical test theory and raw scores.

Magdalena Jelonek (magdalena.jelonek@uek.krakow.pl)
Faculty of Economics and International Relations, Cracow University of Economics, Krakow

Inverse propensity scores as weights in educational research

Keywords: propensity scores, quasi-experimental techniques, regression

This presentation reviews problems of working with observational data and making causal inferences using propensity scores as weights. It discusses and illustrates how propensity scores are used to improve inferences with quasi-experimental data. This presentation is based on the counterfactual model of Rubin (1974, 2010), Shadish (2010) and the Shadish, Cook, Campbell (2002) work on the main threats to internal validity. In observational studies - opposite to the experiments where each subject is randomly assigned to a treated group or a control group - characteristics of the intervention group do not start of the same as that of comparison group. In such a situation propensity scores may be used to improve the validity of estimates. PS can be used in a variety of ways in quasi-experimental research: to stratify, match samples or weight data for balancing main control and experimental groups characteristics. The author in her presentation focus on the last of the approaches. To illustrate these methodological problems author use the data from Study of Human Capital in Poland (SoHCiP) (students survey and fields of study data) on public intervention in the form of commissioned fields of study. The research question refers directly to the effect of the intervention (the purpose of intervention was to increase the number of graduates of fields of study of crucial importance for knowledge-based economy). The conclusion is that this effect was relatively

small, when we consider the value of financing. Additionally, we can formulate the question whether the observed effect of growing number of students of the strategic fields of study have to be caused by the intervention itself.

Presentations in Session E4:

Christoph Helm (christoph.helm@jku.at)
Department for Education and Educational Psychology, Johannes Kepler University, Linz

Computer Adaptive Testing in the Vocational Training Domain 'Accountancy'
Keywords: Accountancy, Competence Modelling, Computer-based Adaptive Testing, Vocational Education

Austria is one of the OECD countries with the largest share of students in vocationally-oriented upper secondary education (OECD, 2012). These vocational fulltime-schools aim to help students to attain A-levels in commercial areas, especially in Accountancy. However, the few existing approaches to develop psychometric tests that allow assessing students' competence in the domain Accountancy are designed for use in the German dual training system (Winther, 2010), and thus not appropriate to diagnose Austrian students' competence at upper secondary school. Given the lack of appropriate tests the so-called WBB (Helm, 2014) was constructed – based on ECD and IRT framework – for assessing students' academic achievement at grade 9 through 11 (using vertical scaling methods; Kolen and Brennan, 2004). However, like all paper-pencil-tests the WBB has several disadvantages like the need for manual scoring and estimation of students' theta, common test length for all students etc. Thus a computer-based adaptive test (CAT) was developed together with researchers from the Institute for Informatics at JKU Linz. The main objective of the presented paper is to show how this CAT was developed by referring to essential fundamentals from content-related didactics (domain and instructional analysis), psychometrics (IRT and CAT background) and informatics (java-implementation). Based on a sample of approx. 1000 students that did the paper-pencil version of the WBB item parameters were estimated against the background of the Rasch model. These empirical results were used to create an item-bank that represents the basis of the catR-algorithm in R (R Core Team, 2014). A first (German-speaking) beta-version of the CAT can be investigated on http://adaptivetesting. ce.jku.at/. Since the test is still under construction validation studies are the next step in this project. One goal of this test development is to provide students and teachers with diagnostic information on their skills in Accountancy. Whereas so far students' could login on the website www.edumetrics.at to get feedback on their test performance, the CAT automatically informs students about which tasks they solved correctly and which they did not. This kind of diagnostic information is discussed against the background of Cognitive Diagnosis Models in vocational education (Helm, Trost, George, and Pocrnja, in press).

Elgars Felcis (elgars.felcis@uj.edu.pl)
Institute of Sociology, Jagiellonian University, Krakow

Issues with item response theory (IRT) and computerized adaptive testing (CAT) practical application in the Baltic States

In spite of decades of IRT and CAT development in various fields, substantial differences persist regarding to its applications in different areas of the world. Spatial focus of this article is on the three Baltic States – Estonia, Latvia and Lithuania – because of their similar historical experiences, including educational systems. The Baltic States have small populations (1.3, 2 and 3 million) – a key factor of IRT and CAT practical development typically requiring large scale applications. Therefore analysis of these countries aims to suggest the scope and potential for IRT and CAT application in small countries or regions. Up to

now IRT is rarely used in each of the countries and there are no examples of CAT use. IRT is typically used by the experts involved in international studies (e.g. PISA) and by the state examination centres in assessment of national school examination result analysis, but not at the stages of examination development thus disabling the use of CAT. Since secondary education provides the largest scope for IRT and CAT application, the accumulation of data over years of examinations could be the basis for item bank creation for different national school examinations. However, multiple aspects of examination preparation and administration in the Baltic States have been well developed without IRT or CAT application and would not provide any additional benefits.

Mateusz Blukacz (mateuszblukacz@gmail.com)
Institute of Psychology, Jagiellonian University, Krakow
Marta Kromka
Institute of Pedagogy, Jagiellonian University, Krakow

Learning strategies as a predictor of academic achievement in Polish 15-year-olds
Keywords: academic achievements, cultural context, learning strategies, PISA

All students learn, but different ways of learning and controlling the acquisition of information/knowledge may be preferred. Learning strategies refer to various mental processes which affect the way the knowledge is gathered and how efficiently new information are assimilated. OECD's Programme for International Student Assessment (PISA) distinguishes three learning strategies: memorization, transfer through elaboration, and metacognition. It has been proven that the relative effectiveness of different learning strategies is diverse and depends on the type of knowledge being processed. Chiu, Chow and Mcbride-Chang (2007) created a comprehensive multilevel model explaining adolescent scores in mathematics, science, and reading achievement by accounting learning strategies in PISA 2000 across 34 countries. Apart from individually used learning strategies, a macro-level environmental components such as countries' economy, cultural context and ecological explanatory factors were taken into account. The aim of this study is to provide an overview of the specific profile and effectiveness of learning strategies used by Polish students by recreating the model on PISA 2009 data. Knowing and understanding the pattern of students' learning strategies may be helpful for both students to learn efficiently and for teachers to select appropriate teaching strategies, improving weaker points, informing the teaching and learning process to enhance achievement and inclusion.

Daniela Wetzelhütter (daniela.wetzelhuetter@jku.at), Johann Bacher
Department of Empirical Social Research, Johannes Kepler University, Linz

How to Measure Participation of Pupils at School Analysis of Unfolding Data Based on Hart's Ladder of Participation
Keywords: CATPCA, Ladder of Participation, MDS, Participation-Degree, Unfolding

In recent years a renaissance of the scientific focus on participation in childhood, youth and educational research has taken place. Research in this context often refers to the "Ladder of Participation" developed by Hart (1992). But, little attention has been paid to its measurement. Therefore, this paper proposes and analyses a measurement instrument for the frequently cited "Ladder of Participation". Hart's (1992) model assumes that eight different levels (degrees) of participation exist, whereby the extent of participation increases with each level (stage) from a low (pretended participation) to a high (supported self-responsibility) degree of participation. Unfortunately, he does not clearly define the dimensional structure of his concept.

Therefore, the paper addresses the following research question: Can participation in school be measured uni-dimensionally as assumed by Hart? In order to answer the research question a rating scale was developed and implemented in an Austrian study on school participation. The scale consists of eight items measuring those eight steps of the ladder. Multidimensional scaling and CATPCA (categorical principal component analysis) were performed as suitable analyzing procedures. In contrast to Hart's (1992) "Ladder of participation", we were not able to distinguish eight different degrees of participation. Three groups of participation: "sufficient participation", "symbolic participation" and "deficient participation" were identifiable. In future studies we will try to improve the measurement instrument in order to be able to verify if only three groups of participation are seen by school children.

Zofia Bednarowska (zofia.bednarowska@uj.edu.pl), Mateusz Lichoń
Institute of Sociology, Jagiellonian University, Krakow
Marcin Kautsch
Institute of Public Health, Jagiellonian University, Krakow

First aid education: knowledge measurement among pupils
Keywords: educational results, first aid education, knowledge measurement

In case of an medical emergency, appropriate provision of the first aid significantly increases chances of of victim's survival. It is crucial to ensure that young people have access to first aid courses that not only equip them with theoretical knowledge but also prepare for its application. Key research question is whether and to what extent differences in educational factors including but not limited to teaching systems, sources of knowledge, number of curses, etc. influence pupils' knowledge and attitudes toward first aid. The research also aims to identify if pupils' knowledge influences their motivation to act. Research scope includes international comparison but focuses on Polish sample. The research consists of two studies: one focused on measuring pupils attitudes towards first aid and the second one measuring their knowledge. PAPI and CAWI surveys were conducted in December 2013, January 2014 on a sample of 3196 pupils of Polish and 153 of Swiss secondary schools. Therefore our main focus is on the Polish sample. The substantive results demonstrate consistent answers of both national groups. One of the key international difference is that Polish surveyed pupils present higher motivation to performing first aid. A considerably high indicator of right answers in the test and declared motivation to provide help is not consistent with pupils' self-confidence. Therefore, it fits into the general challenge in Polish education – it teaches better in terms of knowledge rather than in terms of behaviour and attitude.

Presentations in Session L2:

Marcin Kocór (marcin.kocor@uj.edu.pl), Barbara Worek
Center for Evaluation and Analysis of Public Policies, Jagiellonian University, Krakow

Is it worth it to invest in a training? The problems of measuring the benefits of investment in training and personal development

Keywords: human capital, non-formal education effects measurement, skills development, training effects

The studies of human capital development are focused on two methods of skill improvement: formal education and training. However, often studies account the impact of formal education on the individuals' situation and increase of their salary. Therefore, results of such research conducted in many countries correlate higher education with higher salary, better chance of finding a job and in general, better career prospects. Investigation of the effects of trainings on human capital is taken not so often. The current literature on this subject from countries such as the UK, Netherlands, Sweden and Germany show that participation in trainings contributes to the increase in salaries of employees; however no such relationship was established in the case of Denmark and France. Moreover, the size of this effect differs significantly, which highlights the possible difficulties in measuring and assessing training impact on human capital development. Similar studies on the effects of trainings on employment and wages were conducted in Poland by Dębowski and Pogorzelski (2010) using propensity score matching. The results indicate that participation in non-formal education increased probability of finding work for the unemployed and it has a positive effect on the growth of wages and promotion. However, such research also emphasize on the challenges associated with estimating the impact of non-formal education on human capital development. Such research on this subject, therefore, mark our starting point of reflecting on the possibility of evaluating the effects of training, the nature of the data and possible analytical approaches. We will show different approaches to measure the effects of training and general non-formal education on skill improvement by giving some empirical examples.

Diana Turek (diana.turek@uj.edu.pl)
Center for Evaluation and Analysis of Public Policies, Jagiellonian University, Krakow
Anna Anzulewicz
Institute of Psychology, Jagiellonian University, Krakow
Barbara Ostafińska-Molik
Institute of Pedagogy, Jagiellonian University, Krakow

Exploring mismatch between workers skills and job requirements in selected European countries

Key words: education, labour market, PIAAC, skills mismatch

Skills mismatch on the labour market has become a substantial concern among policy makers. A rich literature defines skills mismatch as the discrepancy between the skills of employed workers and the requirements of the jobs that they occupy (e.g. Quintini, 2011; CEDEFOP, 2010). Mismatch can be described as situation in which workers' skills exceed or lag behind those employers seek (Handel, 2003). The aim of our study is to explore the problem of skills mismatch in selected European countries. The three main research questions of our analysis are: (1) To what extent adults' skills meet job requirements? (2) Do

socio-demographic and job characteristics influence the likelihood of skills mismatch? (3) What is the overlap of mismatch across skills (i.e. literacy, numeracy and problem solving)? Our study was based on Survey of adult skills (PIAAC) data. The major advantage of PIAAC is the availability of measures of proficiency in three important skill domains, namely numeracy, literacy and problem solving. The background questionnaire of PIAAC also includes a very detailed section about the use of skills at work. To measure skills mismatch, we used the framework proposed by Fichen and Pellizzari (2013) in which workers are classified as well matched, over-skilled or under-skilled. In order to answer the research questions we used regression analysis. We compared Polish adults with their counterparts in selected European countries. Negative consequences of skills mismatch on the labour market can be seen at various levels, e.g. individual, company, as well as macroeconomic level. Skills mismatch affects job satisfaction and wages, but also reduces productivity of firms and GPD growth due to waste of human capital. The better understanding of different types of skills mismatch is essential to limit negative outcomes and to design an appropriate policy strategy. A better match of the employees' skills to the requirements of their jobs will not only lead to important improvements in individual workers' well-being, but will also have serious economic impact.

Szymon Czarnik (szymon.czarnik@uj.edu.pl)
Institute of Sociology, Jagiellonian University, Krakow
Krzysztof Kasparek
Center for Evaluation and Analysis of Public Policies, Jagiellonian University, Krakow

Does education work differently for males and females in the labour market?

One of the most spectacular socio-economic changes of the last two decades in Poland has been a sharp rise in the percentage of people with an academic degree (from 14% in 2002 to 40,5% in 2013 for persons in their early thirties). The growth has been particularly dynamic for women. However, fields of their studies have been quite different from those of men. As education level is one of the crucial factors affecting person's situation in labour market, we set out to investigate whether this factor works differently for males and females. The main objective of the present study is to investigate male-female differences in patterns of economic activity, controlled for level of education. One aspect of the problem is to scrutinize the link between shifting education levels and kinds of occupations available for persons with different educational background. We use contingency ratios to measure how various education levels affect probabilities of working part- or full-time, being unemployed, or economically inactive. Probabilities of working in particular occupations, conditional on age, sex and education level, are estimated by means of multinomial logistic regression. All analysis were conducted on the Study of Social Capital (Bilans Kapitału Ludzkiego) data bases. Data were collected yearly from 2010 to 2014 and each year's random sample size was 17,600. The changes in likelihood of having a job brought about by different education levels are similar for males and females. However, secondary or higher education provides a better protection against unemployment for men than women aged 25-40. We consider whether the observed differences are due to biological factors associated with procreation, or due to varied job opportunities offered by different fields of studies.