KoKoHs

Modeling and Measuring
Competencies in Higher Education
Http://www.kompetenzen-im-hochschulsektor.de/index_ENG.php

KoKoHs Working Papers

No. 7

Sebastian Brückner, Simone Dunekacke & Roland Happ

Causal Analysis Using International Data

Report from the "AERA Institute on Statistical Analysis for Education Policy" from 6th till 9th May 2014 in Washington, DC.

SPONSORED BY THE

Federal Ministry
of Education
and Research

Johannes Gutenberg University Mainz

Humboldt University of Berlin

***KoKoHs Working Papers* from the BMBF-funded research initiative „Modeling and Measuring Competencies in Higher Education“**

The *KoKoHs Working Papers* series publishes articles from the funding initiative „Modeling and Measuring Competencies in Higher Education (KoKoHs)“. These may be conceptual papers or preliminary results intended for rapid dissemination or public discussion. Publication in *KoKoHs Working Papers* does not preclude publication of the texts elsewhere. The responsibility for the content lies with the authors. The content does not necessarily reflect the views of the publishers of *KoKoHs Working Papers*.

Contact:

miriam.toepper@uni-mainz.de
corinna.lautenbach@hu-berlin.de

The research initiative is funded by the German Federal Ministry of Education and Research under grant no. 01PK11100A and 01PK11100B.

The *KoKoHs Working Papers* are also available for download:

http://www.kompetenzen-im-hochschulsektor.de/index_ENG.php

Causal Analysis Using International Data

Report from the "AERA Institute on Statistical Analysis for Education Policy" from 6th till 9th May 2014 in Washington, DC.

Brückner, S., Dunekacke, S. & Happ, R.

Contact:

miriam.toepper@uni-mainz.de
corinna.lautenbach@hu-berlin.de

Bibliographical references:

Brückner, S., Dunekacke, S. & Happ, R. (2014). Causal Analysis Using International Data – Report from the "AERA Institute on Statistical Analysis for Education Policy" (KoKoHs Working Papers, 7). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Causal Analysis Using International Data – Report from the "AERA Institute on Statistical Analysis for Education Policy"

Abstract:
International comparative studies in the school sector have been made popular especially by research projects in PISA and TIMSS and have attracted the attention of the general public. The findings were discussed controversially in the involved countries and provide the basis for political decisions in educational matters. However, when executing and evaluating international comparative studies like these, one must proceed with caution, in order to not come to wrong conclusions that may have negative consequences. In this working paper, the results of the workshop on methods led by the AERA Institute on Statistical Analysis for Education Policy are presented in May 2014 in Washington, DC. After a short introduction and presentation of the research projects PISA and TIMSS, which serve as underlying data for the statistical modeling, selected statistical challenges that were theoretically introduced to and practically applied by the workshop's participants will be outlined. The results indicate that while there is great potential in international comparative educational studies, there are some risks to be considered during the actual implication and evaluation of data sets of this sort.

Keywords:
International Comparisons, PISA, TIMSS, Causal Analysis

Contents

1 Introduction

The findings of international comparative studies like ‚Trends in International Mathematics and Science Study (TIMSS)‘ (Chudgar, Luschei & Fagioli, 2012) and ‚Program for International Student Assessment (PISA)‘ (Fleischman, Hopstock, Pelczar, Shelley & Xie, 2010) have turned the focus of the general public to international comparative studies and have led to lively discussions about the respective educational systems of the participating countries. Extensive reports were published in the national media of the participating countries that show how great the public interest in the results of these international comparative studies was and still is (OECD 2002). There has been a sufficient coverage on the advantages of such international comparative studies (Chmielewski, Dumont & Trautwein, 2013).

There are, however, also some critical voices that particularly refer to the risks behind these studies (Schmidt, Zoido & Cogan, 2013). These critical voices refer to the planning and implementation of such studies as well as especially the evaluation of the gained data sets. A very critical approach to the results of these studies is a necessity that should be demonstrated at all times (Schmidt, McKnight, Cogan, Jakwerth & Houang, 1999). This has not only been discussed in the international context, but has also nationally been pointed out by some scientists in Germany. This raises the question how studies like these can be used purposefully to develop and improve the educational systems of the respective countries. Since the beginning of the studies there have been lively discussions about the conclusions drawn concerning the transformation of the educational systems (Terhart, 2002).

The AERA Institute on Statistical Analysis periodically offers method workshops that address specific methodological questions from the field of educational science. From May 6th to 9th 2014 a workshop on the topic ‘Causal Analysis Using International Data’ took place in Washington, D.C. Notable scientists involved in the studies TIMSS and PISA presented selected methodological challenges especially with a view to the analysis of international comparative data sets. The speakers at the workshop include:

- **William Schmidt (Michigan State University)**

  William H. Schmidt is a University Distinguished Professor at Michigan State University and director of the Center for the Study of Curriculum. He serves as co-director of the Education Policy Center and holds faculty appointments in Statistics and Education. Previously he served as National Research Coordinator and Executive Director of the US National Center which oversaw participation of the United States in the IEA Third International Mathematics and Science Study (TIMSS).

- **Martin Carnoy (Stanford University)**

  Martin Carnoy is Vida Jacks Professor of Education and Economics at Standford University. He is a fellow of the National Academy of Education and of the International Academy of Education. He served six years on the Advisory board of Mexico's National Institute of Educational Evaluation.

- **George Wimberly (American Educational Research Association)**

  George L. Wimberly is the Director of Social Justice and Professional Development at the American Educational Research Association (AERA). He is the co-principal investigator on the National Science Foundation funded project, Advancing Knowledge and Building the Research Infrastructure in Education and STEM Learning.

- **Felice Levine (American Educational Research Association)**

  Felice J. Levine is Executive Director of the American Educational Research Association. Dr. Levine is a member of the National Research Council Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioural and Social Sciences.

- **Richard Houang (Michigan State University)**

  Richard T. Houang is Director of Data and Research, Center for the Study of Curriculum at MSU and holds faculty appointment in measurement and quantitative methods.

- **Amita Chudgar (Michigan State University)**

  Amita Chudgar is an Associate Professor of Educational Administration and Education Policy. Her work examines the influence of home, school, and community contexts on educational access and achievement of children in resource-constrained environments.

This paper intends to highlight selected contents from the workshop. Of course it is impossible to present the entire content of the four-day-long workshop in detail. For this reason, we have added references for further information on the topics. Three junior researchers from the KoKoHs funding initiative (Sebastian Brückner, Simone Dunekacke and Roland Happ) were able to take part in this workshop. Apart from lectures on the specific topics, the 33 participants of the workshop were also confronted with practical contents by means of carrying out their own analyses based on provided TIMSS and PISA data sets. We want to thank the AERA Institute on Statistical Analysis for Education Policy for the opportunity to take part in this workshop!

2 Overview on the Studies TIMSS and PISA

In order to make the workshop as practically relevant as possible and to have the participants create practical statistical models, the methodological challenges presented in chapter 3 are implemented based on TIMSS and PISA data from various waves of these studies. To ensure that participants of the workshop as well as the readers of this paper can understand the carried out analyses, we will begin by presenting essential fundamentals of these studies and the corresponding basic works. This presentation does not cover the complete contents of the studies, but aims to point out central problem areas of these data sets that must be considered during the statistical modeling. In order to get a deeper understanding of these studies, we recommend the basic works that were to be prepared for the workshop. These are the source materials:

„Program for International Student Assessment (PISA)“

International References:

Organization for Economic Cooperation and Development. (2012). *Program for International Student Assessment PISA 2009 – Technical Report.* Paris: OECD.

National References:

Stanat, P., Rauch, D. & Segeritz, M. (2010). “Schülerinnen und Schüler mit Migrationshintergrund.” In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 200-230). Münster, Germany: Waxmann.

„Trends in International Mathematics and Science Study (TIMSS)“

International References:

Mullis, I. V. S., Martin, M. O., Foy P. & Arora A. (2012). *TIMSS 2011 International Results in Mathematics.* Chestnut Hill, Mass.: TIMSS & PIRLS International Study Center, Boston College.

National References:

Bos, W., Bonsen, M., Kummer, N., Lintorf, K. & Frey, K. (Eds.) (2009). *TIMSS 2007. Dokumentation der Erhebungsinstrumente zu Trends in International Mathematics and Science Study.* Waxmann: Münster u.a.

Significant differences in the survey design and in the gathered constructs between TIMSS and PISA are summarized in the following table 1. These differences are of interest in two ways. On the one hand, the scientist should consider at the beginning which specific question he wishes to pursue and which data is suitable in this context. On the other hand, the underlying differences in the design are also of importance when combining the data of these two studies (see chapter 4).

TIMSS	**PISA**
Grade level	Age
Classrooms	Schools
Knowledge	Literacy

Tab. 1 Differences in the design between TIMSS and PISA

It becomes obvious that there are considerable differences in the design and especially in the sampling (classroom vs. schools) between the two studies. TIMSS samples 1-2 classrooms in each school. PISA samples students across classrooms within a school. The examined construct differs between the two studies as well. The focus in TIMSS lies on the content knowledge of mathematics and the focus is in the curriculum in mathematics. PISA, on the other hand, aims at literacy and specifically „applied mathematics literacy". These differences must be considered in the statistical model and therefore also evaluated critically at the beginning of the analyses. The conclusion of this comparison is therefore **„Different Design, Different Possibilities for Analyses"** (Schmidt, 2014a).

When presenting the studies, considerable challenges in the data sets are pointed out time after time. 41 countries took part in TIMSS (1995). This study incorporated textbook analyses, standards analyses, a teacher questionnaire, and a student questionnaire. Students in grades 3, 4, 7, 8, and 12 were sampled. Even video studies were used to some extent. In the subsequent TIMSS studies (1999, 2003, 2007, 2011), a student questionnaire, a teacher questionnaire, and a school questionnaire were implemented as well. As opposed to TIMSS, the PISA study aimed at different main focuses in literacy. These main focuses were different depending on the year of the PISA data collection. PISA focused on mathematics in 2003, science in 2006, language in 2009, and mathematics again in 2012. The survey consisted of a school questionnaire, a parent questionnaire, and a student questionnaire.

Questions concerning the design of international comparative studies should be reflected critically. Especially when one wishes to combine data from both surveys, which can be an added value from a methodical and textual point of view, these facts must be considered critically. The TIMSS test was used on graduates of the 12th grade. These graduates, however, must be characterized very differently depending on the country (age etc.). Based on these considerations, one should critically reflect whether the focus should be on the age of the students or rather the entire year group.

From the practical experience of **William Schmidt (Michigan State University)** and **Martin Carnoy (Stanford University)** specific substantial suggestions for international comparative educational studies are highlighted in the workshop:

- **Advice 1:**
  - **As a scientist, never analyze countries when you are not also practically familiar with their structures.**
  - Even when you received detailed information from the countries' ministries, there can be enormous differences in the practical implementation of these curricular specifications.
  - Schmidt expresses a plea to form cross-national research groups in order to combine practical experiences with the countries' educational systems.

- **Advice 2:**
  - **Be careful when analyzing secondary data.**
  - All analyses concerning TIMSS and PISA can be characterized as analyses from secondary data.
  - It is urgent that you think as much about data collection here as you do when working with primary data. At first sight, this necessity does not always appear as such. Primary data of course offers the researcher a much better view into the design of the questionnaire, selection of the sample, limitations when collecting data. For this reason primary data is preferred here. If, however, primary data concerning certain questions cannot be acquired, important steps like missing data, coding of items, direction of scales, linkage of teacher to student etc. must be critically and extensively examined.

3 Structure and contents of the young researchers' sessions

3.1 The role of bias in international comparisons

Researchers using data from international educational assessment and combining this data with contextual data from other forms of assessment run the risk of drawing wrong conclusions if they do not adequately consider the sociocultural background in comparative studies (Buckley, 2009). Unfortunately, evidence increasingly suggests that many observed cross-national or cross-cultural differences are, in fact, contaminated by artifacts of measurement (Johnson, 2003; Rossi, de Vijver & Leung, 1997; Poortinga, 1989). These artifacts can be caused by the test itself (1), the collected sample (2) or insufficient construct specifications (3):

(1) When using different closed answer formats (e.g. likert scales, multiple choice questions), cross-cultural differences in how test persons tick off boxes become obvious. Chinese students tend to place their check mark in the middle of scales, whereas US American students tend to be more positive in their responses (Buckley, 2009). This different perception of items can lead to bias in the analyses, regardless of the content of a question.

(2) Aside from this formal criterion of the test format that can affect international comparability, bias can also be caused when drawing the sample. During low-stakes tests as opposed to high-stakes tests, for example, test persons are considerably less inclined to take part and to properly answer the questions, seeing as the successful participation in the test is in no way connected with personal consequences (Cole & Osterlind, 2008). During explorative tests in higher education that carry no immediate consequences for the students and which students cannot explicitly prepare for like for tests which take place as part of a course, there is the risk of a bias being created on a hardly randomized sample, which can possibly not adequately show a learning process.

(3) Furthermore, a bias can be created on a textual basis, when answering an item i has the effect of creating a variable z which influences the linear correlation of the two variables x and y. This bias is then documented through a measuring error. For this reason, it is all the more important to record all the variables that could influence the correlation between x and y before a measuring in order to be able to control measuring errors. A typical measuring in educational research is therefore always the difference of a true expected value E(Y) and the measured value E(Yo). Seeing as it is impossible to draw random samples in reality, it is necessary to minimize bias in analyses by adding various other indicators. On the basis of these statistical preliminary considerations, correlations between variables can be analyzed precisely in international comparisons. The main point of interest lies on the analyses of the covariance of different indicators, in order to be able to analyze the bias that inevitably arises when analyzing the influence of a variable x on another variable y.

For this reason, we give the following general advice concerning the combination of data sets with PISA data. These aspects must be considered when comparing countries, in order to be able to draw causal conclusions (Schmidt, 2014a):

- Emphases of the data sets,
- Estimation of plausible values,
- Aggregation of the variables that are of interest during the analysis,
- Differences of standard deviations and standard errors,
- Consideration of the multi-level structure between and within the countries that are to be compared (country, district, school, student).

Seeing as PISA contains an enormous number of variables that provide a variety of information on the students, the teacher or the school in the different countries, the focus of the workshop was placed on the socio-economic status (SES) as well as the opportunities to learn (OTL). These are attributed to have strong explanatory power for the test performance measured in PISA and TIMSS. Essential methodological challenges are therefore demonstrated on the basis of these two scales from the PISA and TIMSS data sets. For this reason, we refer at this point to these two theoretical foundations.

For the matters of KoKoHs there is a clear connection to these topics, because many projects are not only considering to measure competencies in Germany only, but are also trying to compare their results to the findings in several countries (see for example the papers in Shavelson & Zlatkin-Troitschanskaia, 2015). For example the project WiwiKom is focused on development of a valid test for measuring university students' competencies in business and economics for Germany and the comparisons of these results with the results from U.S., Japan and Mexico (for more details see Zlatkin-Troitschanskaia, Förster, Brückner & Happ, 2014). In the analyses it is very important to prove for measurement invariance, because it has to be ensured that in several groups from different countries the same construct is measured (Förster et al., 2015). So the workshop was of great relevance to this international point of view of some of the projects of the KoKoHs initiative.

3.2 Issues related to the measurement and the use of SES

For cross-national analysis of test scores, it is important to consider the socio-economic status. It is possible to conceptually differentiate between different kinds of SES that influence the gradual manifestation of test forms. These include the cultural capital, the human capital, the economic capital as well as the social capital (Carnoy & Rothstein, 2013). In terms of the cultural capital, the different levels of background of the students when they start school must be analyzed. It is understood that students from families with a higher level of education are better prepared for school, especially concerning language requirements and the relationship between adults and children (Buchmann & Hannum, 2001). The human capital marks the academic and economic expectations of families as well as the time and effort that families put into their children's education. These are significantly more pronounced in families with a higher level of education. Likewise, financial resources prove to be a strong influence on the possibility of access to the sector of higher education, also referred to as economic capital. A reinforcing contextual effect can also be seen in families with more economic resources living with families that support learning and the informal settings in a similar way. This is referred to as social capital (Chudgar, Luschei & Fagioli, 2012).

In multi-level analyses it was shown on a cross-national level that while these effects influence the mathematics test scores in different countries significantly, they are considerably more pronounced in wealthy countries (Schiller, Khmelkov & Wang, 2002). The phenomenon of family background has so far scarcely been realized with hierarchical models in international comparative studies (Chudgar, Luschei & Fagioli, 2012). Apart from hierarchical models, the influence of the socio-economic status can also be implemented using the classic OLS regression. The conceptual foundations of these methodical aspects were deepened in the workshop.

In order to systematically understand the concept of SES, from a methodical point of view the question arises, which indicators should be considered for this purpose. So far, the indicator "number of books in a family" has proven to be clearly positive and more important than other indicators like „number of articles in a household" or even „family income". All these indicators, however, appear to be highly correlated and to have a strong variation in different countries. This must definitely be considered for statistical modeling. For example, families in the United States own more physical articles per household than families in Korea. Therefore the US families have a higher status.

For international comparisons it should be noted that PISA deviates from earlier approaches of SES data collection and has introduced their own index, the „Economic, Social and Cultural Status (ESCS) Index", which is operationalized by various other criteria like the father's employment, the mother's educational background, the articles in the household or the number of books in the household. Due to its lacking international comparability and its lacking precision concerning the causes for an explanation of the differences in test performances, the index has not been used on multiple occasions (Hauser, 2013; Ehmke & Baumert, 2007). Especially when comparing identical questions with identical test persons in the studies TIMSS and PISA, significant differences became obvious, so that the reliability of the collection of data on the socio-economic status must be questioned (Carnoy, 2014b). In practical exercises, the influence of the factor ESCS on students' performance in the PISA studies was examined. The leader of the workshop was glad to assist the participants of the workshop with any questions. Depending on their preference, the participants were offered to make their analyses on the basis of SPSS, STATA or SAS.

In the German higher education sector many groups of students from different countries are enrolled in the same courses of study. Thus, there is a high heterogeneity of students in the German higher education sector with many different social backgrounds. These different aspects should be taken into account by a comprehensive assessment of the student's socio-economic status. Therefore a big challenge for KoKoHs is also to include these aspects in their assessment designs in different domains in the higher education sector. First evidences in the domain of business and economics show that the socio-economic status of students has an influence on the acquisition of economic and business competences (e.g. students with another mother tongue than German seem to be handicapped) (Happ, Schmidt & Zlatkin-Troitschanskaia, 2014).

3.3 The role of opportunity to learn (OTL)

In order to determine a causal effect, the researcher must control as many variables as possible that might, aside from the anticipated independent variable, influence the dependent variable from a theoretical point of view. All these variables, as a rule, cannot be controlled (Schneider, Carnoy, Kilpatrick, Schmidt & Shavelson, 2005, pp. 16ff). Aside from the sociodemographic variables as described in chapter 3.2, it is a common practice to also collect data on the OTL in educational scientific studies that analyze the outcome of educational systems (Schmidt, Zoido & Cogan, 2013, pp. 4). Especially when working with international comparative data, the control of opportunities is significant, because there are country-specific differences between the curricula and the OTL (Schmidt, 2014b).

In the early 1960s, OTL were already considered an important aspect of learning. In these first conceptions, particularly the time spent actively with the subject of study was considered a measurement of the opportunities to learn (Caroll, 1963). Later on, apart from this time-related consideration, textual aspects were also included (Schmidt et al., 2001, p. 342). Nowadays, data on OTL is usually collected on student as well as training level in all larger educational systems (e.g. PISA, TIMSS, TEDS-M). In the course of this it becomes clear that this variable serves as an essential control variable between the educational systems and the learning opportunities perceived there. Within the framework of the workshop, it is underlined that it can be deemed necessary for the researcher to create this common standard for comparison. With this, substantial interpretations with a view to the gathered performance data (mathematical capabilities of the test persons) can be formulated.

For the theoretical modeling of the OTL, a model derived from educational science is taken as a basis. During the workshop, a superordinate model was presented that differentiates between multiple levels of OTL (Schmidt, 2014b). To begin with, the intended curriculum is taken up, which encompasses government-mandated goals and standards. The next level consists of the implemented curriculum. A distinction is made here between the potentially implemented curriculum as it is predefined in textbooks and could *potentially* be taught, and the implemented curriculum, that is what *actually* happens in the classroom (aims, methods, etc.). On the third level the attained curriculum is located, which records what the students actually learned.

Four levels can be deduced from this, levels on which OTL can be examined. On the first level, this is an enquiry oriented on the standards predetermined by the system, whether these were covered on class or not. According to the experts present at the workshop, this is fairly easy to put into practice.

Moreover, it is also possible to acquire data on OTL on the level of the potentially implemented curriculum, where teachers are systematically questioned on textbook topics. This approach, however, has proven according to the experts as too detailed and extensive. This can be attributed to the fact that as a rule different textbooks are used and an analysis on class level would require high resources (Schmidt et al., 2001, p.336). A third possibility is the gathering of data on the level of the implemented curriculum. Here, the topics and methods used by teachers in class are gathered. This can optionally also happen by means of the teaching time used. On the last level, the acquisition of data on the level of students is located. Here the question is explored, which contents are necessary in order to work on the tasks or whether they are familiar with them.

In TIMSS data concerning the OTL has been collected for quite a while. OTL are measured by means of teacher questionnaires that are then assigned to the respective classes (Schmidt et al., 2001, p. 349). Three aspects are covered: "1) the mathematics content itself (topic coverage-yes/ no), 2) instructional time for each topic, and 3) rigor or content difficulty (as estimated from international curriculum data)" (Schmidt, 2009, p. 13). This approach can be assigned to the level of the implemented curriculum. In PISA, data on OTL was first collected in 2012 (Schmidt, Zoido & Cogan, 2013). Because of the different conceptualization from TIMSS, that is sampling students not by class but by age, it is not possible to gather data on OTL in PISA by class. For this reason, voluntary information by the students was collected (Schmidt, Zoido & Cogan, 2013). On the one hand, the sample assessed how familiar the students are with the respective topics and, on the other hand, by means of concrete sample items, the students were asked how often the respective problem is addressed in class or in exams (Schmidt, 2014b). Both scales were correlated with one another by Schmidt, Zoido & Cogan (2013), who were able to show that even the students' voluntary information provide valid results (ibid.).

Through hierarchical linear modeling or regressions, the correlation between opportunities to learn and achievements can be worked out methodically, while at the same time considering the clustered structure of the data (students, classes, schools, optionally country). Examples for this can be found in the works of Schmidt, Förster & Zlatkin-Troitschanskaia (2014), Schmidt (2009) and Schmidt et al. (2001) (based on TIMSS data), as well as Schmidt, Zoido & Cogan (2013; based on PISA data).

3.4 The role of variance in cross country comparisons

Aside from the average value, variation and variance are considered an essential criterion for describing the distribution of data. The statistical analysis aims to explain the variance as much as possible. The explanation of the variance is calculated by means of a regression model by asking what percentage of the variance of Y is explained by X. Especially pedagogical studies address the question which part of Y (e.g. the students' performance) the levels that determine the educational system represent. This is a hierarchical structure. As a result, there is on principle a multi-layered structure. The most basal structure results from the consideration of at least students within classrooms as well as classrooms within schools. Furthermore, if necessary, additional levels like schools within districts or districts within countries must be considered. The question arises how the data changes on different levels. In this context, this is referred to as decomposing.

Decomposing occurs on the level of test scores as well as on the level of variance. In this manner, the test score is formed by the total average value as well as the district, school and class effects and the individual difference. Consequently, the variance consists of the variance of the district means, the average variance of the school means within districts, the class means within schools and the individuals within classes (Houang, 2014).

If by this means sources for differences on the various levels are to be identified, a randomized sample on all levels is required, otherwise the estimation of the variance components would be confounded. The individual components for the test score and the variance on the level of districts, schools, classes and individuals can only be determined when at least two units are sampled on the same level. This consequently means that at least two schools from the same district or two classes from the same school must be sampled.

A variance analysis of this kind can be carried out e.g. by means of an ANOVA. Estimation procedures are often already implemented in software. One example is the Maximum Likelihood Estimation, which is implemented in SAS, STATA or MPlus. Moreover, structural equation models can be useful analysis technology, seeing as they allow the modeling of more complex correlations between the independent and the dependent variables (Schmidt, 2014d). A targeted possibility of controlling the explanation of variance is using so-called panel data. During the workshop, the fixed effects models (Carnoy, 2014a) were presented, that allow the consideration of the individual and time-consistent heterogeneity of the individual.

The research program KoKoHs is going to assess students` competencies in different higher education institutions in Germany and some other countries. In order to get a high sample size, like in schools, the students have to be assessed in different courses. In general, there is a hierarchical structure in the higher education sector, in which students are nested in courses, courses are nested in universities which are nested in countries (s. Schmidt et al. 2014). So the methods taking into account hierarchical structures in the school level are generally also applicable to the higher education sector.

4 Discussion and Outlook

International comparative research has great potential to identify cause-and-effect relationships, which can contribute to an optimization of education processes by analyzing data from various countries. With all these advantages, however, problem areas and risks which are particularly related to the different specifications of the examined educational systems are often disregarded. These two perspectives on comparative studies, of advantages but also risks, were repeatedly stressed during the course of the AERA workshop in May 2014. The speakers (see chapter 1 of this article) were able to revert to extensive experience in the field of international comparative research, which they willingly shared with the participants of the workshop.

Although the workshop primarily focused on data from the school sector (TIMSS and PISA), these studies in fact served as a starting point to demonstrate practical experiences in data analysis to the participants. It is obvious that these methodical deliberations, which in particular concern the planning of the design, the evaluation as well as the analysis of the international comparative studies, can also be applied to higher education. Like in the school sector, in higher education similar curricular contents between countries are identified and measuring instruments are constructed with regard to these benchmarks, on the basis of which empirical data can be collected.

SES and OTL should also be reasonably considered in higher education. While researchers working on SES can probably resort to the variables already successfully used in TIMSS and PISA, research on OTL is still in at a very early stage (Blömeke, Fellbrich, Müller, Kaiser & Lehmann, 2008). Here, the development of suitable (and internationally comparable) instruments will presumably be of importance in the future.

A fact what was repeatedly emphasized during the course of the workshop is that studies of this sort must be undertaken with great caution. In studies like these, even the greatest methodical proficiency cannot correct a targeted critical reflection of the collected data from the participating countries. The countries' educational systems and, in the case of the KoKoHs funding initiative, their systems of higher education are to be specifically analyzed concerning similarities and especially differences. These differences can distort the results. This phenomenon was presented in the workshop by means of a detailed analysis of the adolescents' opportunities to learn.

Furthermore, from an economic perspective it is impossible to collect all data necessary for the valid acquisition of certain latent attributes in primary research. For this reason, the combination of different data sets is not only relevant for the school sector, but also for the higher education sector, in order to control bias in the estimation of the latent attributes as well as possible. This, however, should always be done with caution because the data usually refers to different contexts (e.g. SES of the test persons) and levels of aggregation (e.g. individuals, universities, countries). An appropriate way of integrating this data is therefore only possible on the respective highest aggregated level, which creates the necessity of explicitly considering the multi-level structure in the analyses.

References

Blömeke, S., Fellbrich, A., Müller, C., Kaiser, A. & Lehmann, R. (2008). Effectiveness of teacher education. State of research, measurement issues and consequences for future studies. *ZDM Mathematics Education, 40,* 719-734.

Buchmann, C. & Hannum, E. (2001). Education and Stratification in Developing Countries: Review of Theories and Empirical Research. *Annual Review of Sociology*, 27(1), 77-102.

Buckley, J. (2009). Cross-National Response Styles in International Educational Assessments: Evidence from PISA 2006. Available: https://edsurveys.rti.org/PISA/documents/Buckley_PISAresponsestyle.pdf (June 2014).

Carnoy, M. (2014a). Students fixed effects with cross section data. Presentation at the AERA Institute on Statistical Analysis for Education Policy. Washington, DC (unpublished presentation).

Carnoy, M. (2014b). Social class in international test analysis. Presentation at the AERA Institute on Statistical Analysis for Education Policy. Washington, DC (unpublished presentation).

Carnoy, M. & Rothstein, R. (2013). WHAT DO INTERNATIONAL TESTS REALLY SHOW ABOUT U.S. STUDENT PERFORMANCE? Washington, DC: Economic Policy Institute. Available: http://www.epi.org/publication/us-student-performance-testing/ (June 2014).

Caroll, J.B. (1963). A model of school learning. *Teaching College Record*, 64(8), 723-733.

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking Effects Depend on Tracking Type: An International Comparison of Mathematics Self-Concept. *American Educational Research Journal*,50(5), 925-957.

Chudgar, A., Luschei, T. F. & Fagioli, L. P. (2012). *Constructing Socio-Economic Status Measures Using the Trends in International Mathematics and Science Study Data.* East Lansing: Michigan State University.

Cole, J. S. & Osterlind, S. J. (2008). Investigating Differences Between Low- and High-Stakes Test Performance on a General Education Exam. *The Journal of General Education*, 57(2), 119-130.

de Vijver, F. J. R. V. & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga & J. Pandey (Eds.), *Handbook of Cross-Cultural Psychology* (pp. 257-300). Volume 1: Theory and Method. Boston, Mass.: Allyn and Bacon.

Ehmke, T. & Baumert, J. (2007). Soziale Herkunft und Kompetenzerwerb: Vergleiche zwischen PISA 2000, 2003 und 2006. [Social background and competency acquisition: Comparisons between PISA 2000, 2003 and 2006] In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Eds.), *PISA 2006: Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006: The results oft he third international comparative study] (pp. 309-335). Münster: Waxmann Verlag.

Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., Shelley, B. E. & Xie H. (2010). Highlights from PISA 2009: Performance of U.S. 15-Year-Old Students in Reading, Mathematics, and Science Literacy in an International Context (NCES 2011-004). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Available: http://nces.ed.gov/pubs2011/2011004.pdf (June 2014).

Förster, M., Zlatkin-Troitschanskaia, O., Brückner, S., Happ, R., Hambelton, R., Walstad, B.W., Asano, T. & Yamaoka, M. (in review). Validating Test Score Interpretations by Comparing the Results of Students from the United States, Japan and Germany on a Test of Economic Knowledge in Higher Education. In S. Blömeke, J.-E. Gustafsson & R. Shavelson (2015) (Eds.), Assessment of Competencies in Higher Education. Topical Issue of the Journal for Psychology.

Happ, R., Schmidt, S. & Zlatkin-Troitschanskaia, O. (2014). Die Entwicklung sprachlicher Kompetenzen bei angehenden Lehrkräften im kaufmännisch-verwaltenden Bereich und der Einfluss auf den Fachwissenserwerb im Studienverlauf in der Domäne Wirtschaft. [Development of prospective business and economics teachers' language competencies over the course of studies and influence on the acquisition of content knowledge in the domain of business and economics] (Submitted).

Hauser, R. (2013). Some Methodological Issues in Cross-National Educational Research - Quality and Equity in Student Achievement. *Euramerica*, 43(4), 709-752.

Houang, R. (2014). *The role of variance in cross country comparisons. Presentation at the AERA Institute on Statistical Analysis for Education Policy*. Washington, DC (unpublished presentation).

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, 68, 563-583. OECD (Eds.) (2002). PISA in the News in Germany. Paris: OECD Publications.

Poortinga, Y. P. (1989). Equivalence of Cross-Cultural data: An overview of basic issues. *International Journal of Psychology*, 24, 737-756.

Rutkowski, L., Davier, M. v. & Rutkowski, D. (2014). Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis. Chapman & Hall/CRC statistics in the social and behavioral sciences series. Boca Raton: CRC Press.

Schiller, K. S., Khmelkov, V. T. & Wang, X.-Q. (2002). Economic Development and the Effects of Family Characteristics on Mathematics Achievement. *Journal of Marriage and Family*, 64, 730–742.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H. & Shavelson, R. J. (2005). *Estimating Causal Effects. Using Experimental and Oberservational Designs.* Washington, DC: American Educational Research Association.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H. & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Schmidt, S., Förster, M. & Zlatkin-Troitschanskaia, O. (2014). A multilevel analysis of differences in the economic content knowledge of university students in Germany with individual and contextual covariates (American Educational Research Association Congress discussion paper). Philadelphia: AERA.

Schmidt, W. H., McKnight, C., Cogan, L. S., Jakwerth, P. M. & Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education.* Dordrecht/Boston/London: Kluwer.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. C., Wiley, D. E., Cogan, L. S. & Wolfe, R. G. (2001). *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning.* Jossey-Bass: San Francisco.

Schmidt, W. H. (2009). *Exploring the relationship between content coverage and achievement: Unpacking the meaning of tracking in eighth grade mathematics.* East Lansing: Michigan State University, The Education Policy Center.

Schmidt, W. H., Zoido, P. & Cogan, L. (2013). Schooling Matters: Opportunity to Learn in PISA 2012. *OECD Education Working Papers*, No. 95, OECD Publishing.

Schmidt, W. H. (2014). *Causal Modeling Outside the Randomized Experiment. Presentation at the AERA Institute on Statistical Analysis for Education Policy.* Washington, DC (unpublished presentation).

Schmidt, W. H. (2014a). *What we can learn from TIMSS and PISA? Presentation at the AERA Institute on Statistical Analysis for Education Policy.* Washington, DC (unpublished presentation).

Schmidt, W. H. (2014b). *The role of opportunity to learn (OTL) in international comparisons. Presentation at the AERA Institute on Statistical Analysis for Education Policy.* Washington, D.C. (unpublished presentation).

Schmidt, W. H. (2014d). *Multiple Regression and structural modeling. Presentation at the AERA Institute on Statistical Analysis for Education Policy.* Washington, DC (unpublished presentation).

Terhart, E. (2002). Wie können die Ergebnisse von vergleichenden Leistungsstudien systematisch zur Qualitätssicherung in Schulen eingesetzt werden. [How can the results of comparative studies be systematically used for quality assurance in schools] *Zeitschrift für Pädagogik*, 48, 91-110.

Zlatkin-Troitschanskaia, O & Shavelson, R. (2015, in prep.). The international State of the research on assessment of competencies in Higher Education. In Journal Studies in Higher Education (special issue).

Zlatkin-Troitschanskaia, O., Förster, M, Brückner, S. & Happ, R. (2014). Modeling and measuring competencies in business and economics among students and graduates in German Higher Education - Drawing on insights from the German WiWiKom Project. In H. Coates (Ed.), Assessing Learning Outcomes: Perspectives for quality improvement. Frankfurt a.M.: Lang (in preparation).

Previously published:

KoKoHs Working Papers, 1

Blömeke, S. & Zlatkin-Troitschanskaia, O. (2013). Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs [Modeling and Measuring Competencies in Higher Education: Aims, theoretical framework, design, and challenges of the BMBF-funded research program KoKoHs] (KoKoHs Working Papers, 1). Berlin & Mainz: Humboldt University & Johannes Gutenberg-University.

KoKoHs Working Papers, 2

Blömeke, S. (2013). Validierung als Aufgabe im Forschungsprogramm "Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor" [The task of validation in the research program „Modeling and Measuring Competencies in Higher Education"] (KoKoHs Working Papers, 2). Berlin & Mainz: Humboldt University & Johannes Gutenberg-University.

KoKoHs Working Papers, 3

Blömeke, S. & Zlatkin-Troitschanskaia, O. (Eds.) (2013). The German funding initiative "Modeling and Measuring Competencies in Higher Education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students (KoKoHs Working Papers, 3). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

KoKoHs Working Papers, 4

Berger, S., Hammer, S., Hartmann, S., Joachim, C. & Lösch, T. (2013). Causal Inference in Educational Research. Approaches, Assumptions and Limitations. (KoKoHs Working Papers, 4). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

KoKoHs Working Papers, 5

Toepper, M., Zlatkin-Troitschanskaia, O., Kuhn, C., Schmidt, S. & Brückner, S. (2014). Advancement of Young Researchers in the Field of Academic Competency Assessment – Report from the International Colloquium for Young Researchers from November 14-16, 2013 in Mainz (KoKoHs Working Papers, 5). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

KoKoHs Working Paper, 6

Kuhn, C., Toepper, M., & Zlatkin-Troitschanskaia, O. (2014). Current International State and Future Perspectives on Competence Assessment in Higher Education – Report from the KoKoHs Affiliated Group Meeting at the AERA Conference on April 4, 2014 in Philadelphia (USA) (KoKoHs Working Papers, 6). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.