Karlsruher Institut für Technologie

Competence Assessment in Higher Education. A Pilot Study on the Measurement of Competencies in Empirical Social Research Methods among University Students.

von Felix Wolter, Jürgen Schiener und Ingmar Ehler

Nr. 3 | April 2016

WORKING PAPER SERIES IN SOCIOLOGY

KIT – Universität des Landes Baden-Württemberg und nationales Forschungszentrum in der Helmholtz-Gemeinschaft

www.soziologie.kit.edu/socpaper

Impressum

Karlsruher Institut für Technologie (KIT)
Institut für Soziologie, Medien- und Kulturwissenschaften (ISMK)
Schlossbezirk 12
76137 Karlsruhe

Working Paper Series in Sociology

No.3 | April 2016

www.soziologie.kit.edu/socpapers

ISSN: 2363-8079

Competence Assessment in Higher Education. A Pilot Study on the Measurement of Competencies in Empirical Social Research Methods among University Students.

Felix Wolter, Jürgen Schiener, and Ingmar Ehler

Contact:

Dr. Felix Wolter (corresponding author)
Department of Sociology
Johannes Gutenberg University Mainz
Jakob-Welder-Weg 12
D-55128 Mainz
Germany

Phone: +49 6131 39 20831
Email: felix.wolter@uni-mainz.de

Dr. Jürgen Schiener
Department of Sociology
Johannes Gutenberg University Mainz
Jakob-Welder-Weg 12
D-55128 Mainz
Germany

Phone: +49 6131 39 24705
Email: juergen.schiener@uni-mainz.de

Ingmar Ehler
Email: ingmare@online.de

Biographical Information

Dr. Felix Wolter is research assistant in sociology at the Johannes Gutenberg University Mainz, Germany. He was visiting professor at the KIT, Karlsruhe, from 2015 to 2016. His research interests include social-science methodology, statistics, competence assessment, social inequalities, and rational choice theory. Recent publications include „Auf dem Weg zum "Hochschul-PISA"? Zur Messung soziologischer Kompetenzen”, Soziale Welt 65 (1), 2014: 47–73 (together with Jürgen Schiener); „Who Is Telling the Truth? A Validation Study on Determinants of Response Behavior in Surveys“, Public Opinion Quarterly 78 (1), 2014: 126–146 (together with Peter Preisendörfer); and “Asking Sensitive Questions: An Evaluation of the Randomized Response Technique versus Direct Questioning Using Individual Validation Data”, Sociological Methods and Research 42 (3), 2013: 321–353 (together with Peter Preisendörfer).

Dr. Jürgen Schiener is academic director in sociology at the Johannes Gutenberg University Mainz, Germany. His research interests include social-science methodology, statistics, labor market research and sociology of education. Recent publications include „Auf dem Weg zum "Hochschul-PISA"? Zur Messung soziologischer Kompetenzen”, Soziale Welt 65 (1), 2014: 47–73 (together with Felix Wolter); „Weiterbildung im betrieblichen Kontext“, in R. Becker & A. Schulze (Eds.), Bildungskontexte. Strukturelle Voraussetzungen und Ursachen ungleicher Bildungschancen, 2013: 555–594, Wiesbaden: Springer VS (together with Felix Wolter and Ulrike Rudolphi); and „Lehrerurteile und Bildungsgerechtigkeit: Schullaufbahnempfehlungen angehender Lehrkräfte im Vignettenexperiment“, Zeitschrift für Soziologie der Erziehung und Sozialisation 31 (2011): 186–200 (together with Alexander Schulze).

Ingmar Ehler is a sociology student and tutor at the Department of Sociology, Johannes Gutenberg University Mainz.

Competence Assessment in Higher Education. A Pilot Study on the Measurement of Competencies in Empirical Social Research Methods among University Students.

Abstract

In the general education system, standardized competence assessment of students are by now well established (cf. the PISA studies). As concerns the higher education sector, however, evaluation of student achievement is still mostly based on subjective indicators or on indicators that measure input into the educational system (e.g., the funding of universities), although research has clearly pointed out the flaws of this practice. Given the general demand for an objective, standardized competence assessment among university students, it is astonishing that especially in sociology, a discipline which is supposedly sensible for the need of valid measurement instruments and at the same time has the methodological competency to develop them, virtually no research has focused on this issue yet. Our article is intended to start filling this gap. We present results from a pilot study devoted to the definition and measurement of competencies in quantitative empirical social research methods – a core sub-discipline of all social sciences which is particularly well suited for competence measurements – among university students. For this purpose, we present a structural competence model, that was operationalized into test items which were administered to 776 sociology students in Germany and Switzerland. The resulting data were scaled into competence indicators using methods of item response theory. The resulting indicators show satisfactory scale properties and good external validity. Content-related analyses on determinants of student achievement, as measured by the competence indicators, show a fruitful analysis potential of the data. All in all, the results are in favor of further pursuing competence assessments of university students in sociology. For this goal, however, several problems that we also discuss in the paper have to be addressed in the future.

Keywords

Competence assessment, measurement, item response theory, higher education, empirical research methods.

1 Motivation

In the general education system, competence assessments of students are by now well established. In large-scale studies like PISA (OECD, 2014; Prenzel et al., 2008) or TIMSS (Baumert, Bos, & Watermann, 1998; Mullis, Martin, Foy, & Arora, 2012),[1] competencies of students are regularly evaluated on national and international levels with the aid of well-defined competence models and elaborate measurement instruments. Usually, competencies are defined as cognitive dispositions for acting appropriately to domain-specific demands, situations, or problems; they are measured by standardized tests and use methods of item response theory (IRT) for scale construction.

As concerns the higher education sector, however, standardized competence assessment is less advanced. For most disciplines – "domains" in the jargon of the literature on competence assessment – definitions of competencies and measurement instruments are still lacking (Blömeke, Zlatkin-Troitschanskaia, Kuhn, & Fege, 2013). The motivation, however, to fill this gap certainly exists for the same reasons – and additional ones – for which large-scale assessments in the general education sector have been established: First, evaluation in the higher education sector is still mostly based on input-orientated indicators (e. g., facilities and financial situation or teacher-student ratios) and on subjective measures of achievement. Yet, one would prefer an evaluation on the basis of *objective* indicators measuring the *output* of the educational system. In this regard, empirical evidence shows for example that subjective measures yield results that have little in common with those from objective ones, and that the correlations between subjective and objective measures tend to decline with growing objectivity of the measurement instrument (Clayson, 2009; Stehle, Spinath, & Kadmon, 2012). Second, research on determinants and returns of education is mostly based on proxy indicators such as certificates or grades. Here, one would prefer measures of what graduates *really* are able to do, a point that is also of particular importance for the long-standing research tradition on human capital theory (Becker, 1975; Mincer, 1974), which is targeted at the *productivity* of people, but usually relates on proxy indicators in empirical applications. Third, empirical research has shown that grades as the common tool to compare the performance of students are problematic for several reasons; for example, grades especially in the higher education sector are hardly comparable between different institutions (Müller-Benedict & Tsarouha, 2011), which in turn derogates the function of the educational system to assure meritocracy, namely, that individual status should exclusively depend on one's own performance.

Having identified the need and lack of competence models and measurement instruments, some projects have started research into these open issues and evaluate possibilities and limits of establishing competence models and measuring instruments in the field of higher education. Among them are the AHELO[2] project of the OECD (Organisation for Economic Cooperation and Development (OECD), 2013; Tremblay, 2013; Tremblay, Lalancette, & Roseveare, 2012) and over 20 research projects within the framework "Modeling and Measuring Competencies in Higher Education" (Zlatkin-

[1] PISA: Programme for International Student Assessment; TIMSS: Trends in International Mathematics and Science Study.
[2] AHELO: Assessment of Higher Education Learning Outcomes.

Troitschanskaia, Blömeke, Kuhn, & Buchholtz, 2012), funded by the German Federal Ministry of Education and Research.[3]

Despite these developments, little effort has, to our knowledge, been spent on developing concepts and instruments for assessing competencies in the domain of sociology. This is presumably due to the fact that sociology as an academic discipline is unstructured and characterized by different methodological and theoretical paradigms, making it difficult to define and measure "sociological competencies." A first pilot study that we conducted on this issue, however, points to the general feasibility of such projects (Wolter & Schiener, 2014). Yet, one key result of our study was that indeed definition and measurement of competencies in "sociology as a whole" is a complex task – for which reason we have conducted a second pilot study in which we concentrated on a more clearly and narrowly defined, but highly important sub-dimension of sociological competencies, namely, competencies in quantitative empirical social research methods.

In this article, we present results and insights from this second pilot study focusing the development of a scale for the assessment of competencies in quantitative empirical social research methods among sociology students at German speaking universities. This undertaking implicates two main tasks: First, a structural or conceptual model has to be developed in which it is conceptually defined what "competencies in empirical social research methods" stands for. Second, the structural model has to be transferred into test items and administered to a sample of students; the resulting data has, using methods of IRT, to be inspected regarding the scalability and the quality of the resulting scales.

Among all competencies that are conveyed during the formation of sociologists, those in empirical research methods are crucial. For example, Meulemann (2002: 46) points out that given the "pluralism of approaches and the variety of research fields" in sociology, "empirical social research is the fixed point of academic education" [in sociology]. Empirical studies have found that good skills in empirical methods correlate positively with and are highly relevant for chances on the labor market (Schnell, 2002: 38ff.). At the same time and in contrast to other sociological fields, the contents of academic formation in research methods seem to be rather well defined, clearly structured, and more or less consensual among the academic sociological staff, which makes the domain more easily accessible for a standardized assessment of competencies than other domains in sociology. Furthermore, the conscience for the need of valid measurement instruments in general is supposed to be present among empirically working social scientists. At the same time, it is exactly this scientific community, in which the knowledge of methodological techniques required to develop measurement instruments is supposed to be most developed as compared to other experts. Given all this, it is amazing that no research has focused on this issue yet.

In what follows, we will first present key concepts and methods as found in the literature on competence assessment in section 2. Section 3 is devoted to the presentation of a structural competence model, its operationalization into test items, the design of the survey, and the resulting data that forms the basis for the empirical analyses. Empirical findings are presented in section 4, in which we first present the scaling of our competence indicators, followed by selected content-related empirical analyses

[3] For a more detailed overview of ongoing research in these areas, see Wolter & Schiener (2014, 2015).

pointing out the validity of our indicators and illustrating the analysis potential of the data. Our paper concludes with a discussion in section 5.

2 Competence Measurement in Higher Education: Concepts and Methods

Following a definition that a large body of empirical studies is based on, competencies are context-specific cognitive dispositions necessary to successfully cope with certain situations or tasks in specific domains (Klieme & Leutner, 2006: 878f.). Competencies can be learned and acquired by experience in relevant types of situations or by outside intervention. This notion stresses the difference to universal personality traits such as intelligence (which is not conceptualized as learnable) by referring to a certain context, indicating that competencies pertain only to clearly outlined sets of situations and problems and their specific requirements. These situations are assignable to a subject-specific domain which also can be divided into multiple subdomains. Put simply, when talking about competencies we are dealing with “can-do-assertions” or skill descriptions (Pant, Böhme, & Köller, 2012: 50).

A substantial account of a specific competence comprises cognitive processes and objects they apply to and is conceptualized in a structural competence model. The so-called Bloom taxonomy (Bloom, 1956) provides a useful basis for such a model. According to a revised version of this taxonomy (Krathwohl, 2002), cognitive processes such as remember, understand, apply, evaluate and create are arranged in a hierarchy in a way that the more demanding cognitive processes require the less demanding ones.

Competence assessment is necessarily the measurement of competence constructs, which are latent and only indirectly observable through manifest indicators. There are two central criteria that measurement instruments should meet: The level of difficulty of the test items should correspond to the level of ability of the test subjects and the indicators should measure unidimensionally, that is, apart from the difficulty of the test items, only the latent trait of interest should affect the probability of answering an item correctly. This requirement results directly from the definition and the purpose of competence models: If other factors than the intended measured competence affect response behavior, then it is unclear what the items actually measure. Therefore, tests for unidimensionality are inherently validity tests. Statistically this amounts to the requirement for local stochastic independence: When controlling for competence, test items should neither correlate among themselves nor with external variables like gender or the ethnic background of the participants. This requirement is checked as part of the scaling of the competence indicators using methods of IRT.

IRT methods in general aim at modeling a latent variable – the ability or competence to be measured – through empirically observable answers of respondents to test items. The fundamental model which forms the basis for more elaborated ones is the Rasch or 1PL model (Rasch, 1960) which defines the probability of a correct answer on a binary coded item in dependence of personal ability θ and item difficulty δ (formula 1).

$$P\left(X_{ij}=1\right)=\frac{\exp\left(\theta_i-\delta_j\right)}{1+\exp\left(\theta_i-\delta_j\right)} \tag{1}$$

An important characteristic of the Rasch model (and advanced models) is that the ability of the respondents and the difficulty of the items are measured on the same scale. The item characteristic curves (ICC), depicting the probability of a correct response for an item as a function of ability, are parallel in the Rasch specification. The item difficulty is defined as the point on the θ-axis where the probability of a correct answer exceeds a certain value, commonly p=0,5. There is a whole array of model specifications which successively add parameters to the Rasch model in order to achieve a better adaption of the model to the data, but at the cost of abandoning the simple assumptions and some preferable features of the Rasch model (see de Ayala, 2009; Wilson, 2005 for a more detailed discussion). Among these further model specifications, the Birnbaum or 2PL model is frequently used (formula 2). In the Birnbaum model, the assumption of equal discrimination of the items is relaxed and estimated separately for each item. This results in ICCs that may intersect and the rank order of item difficulties to change over the range of θ.

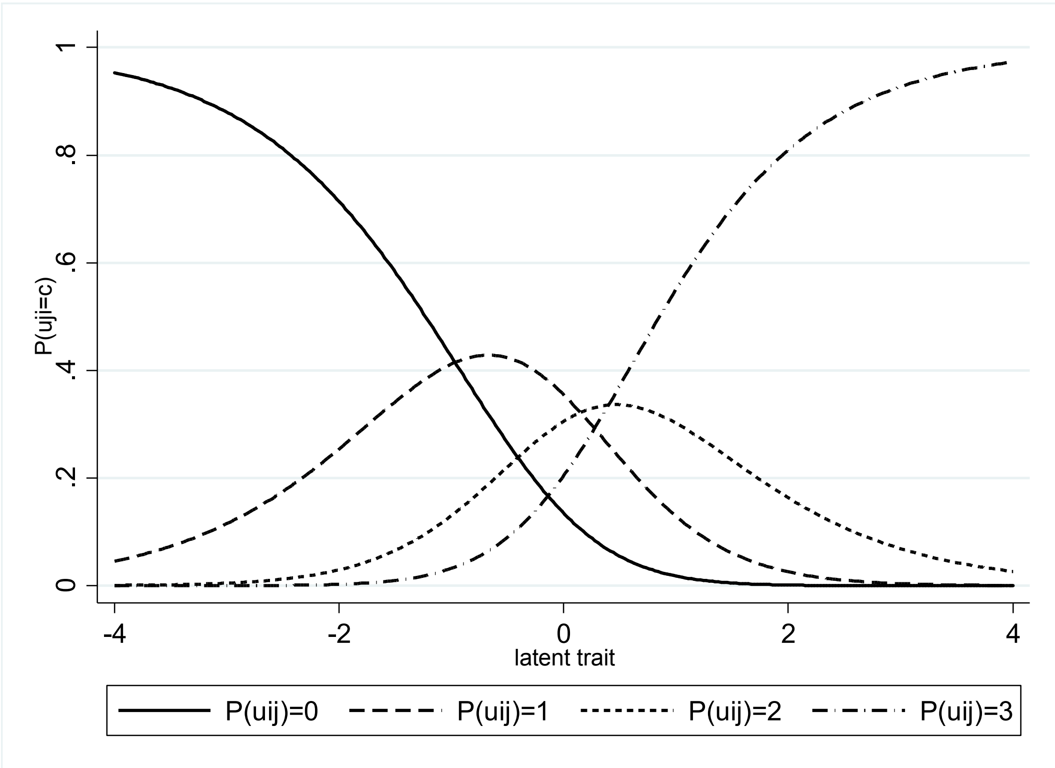
$$P\left(X_{ij}=1\right)=\frac{\exp\left(\alpha_j\left(\theta_i-\delta_j\right)\right)}{1+\exp\left(\alpha_j\left(\theta_i-\delta_j\right)\right)} \tag{2}$$

For test items that encompass not only dichotomous answers (right, wrong), but also partially correct solutions (wrong, partially correct, completely correct), IRT models for ordinal (and nominal) items have been developed. The most important are the partial credit model (Masters, 1982), the general partial credit model (Muraki, 1992), and the graded response model (Samejima, 1969) and its derivatives. Partial credit models (PCM) model the probability of answering the *k*th response category of an item *j* in dependence of person ability θ and the step parameter δ_{j} of the *k*th response category. The respective step parameter refers to the point on the θ scale at which the probability of answering the next higher response category correctly exceeds the probability of a correct answer to the next lower response category (formula 3). Figure 1 depicts the ICC for an example item containing four response categories. Analogously to the Birnbaum model as compared to the Rasch model, the PCM extends to the generalized partial credit model (GPCM) by adding individual slope parameters for each item (Muraki, 1992).

$$P\left(u_{ij} = c \middle| \theta_i, \delta_{jk}, \ldots, \delta_{jm}\right) = \frac{e^{\sum_{k=0}^{c}\left(\theta_i - \delta_{jk}\right)}}{\sum_{l=0}^{m_j} e^{\sum_{k=0}^{l}\left(\theta_i - \delta_{jk}\right)}}$$

$$\sum_{k=0}^{0}\left(\theta_i - \delta_{jk}\right) \equiv 0 \tag{3}$$

Figure 1: Item Characteristic Curves (ICC) in the Partial Credit Model



In order to investigate the psychometric properties of the measurement instruments, various diagnostic tools do exist which can only briefly be mentioned here (see de Ayala, 2009; Moosbrugger, 2012; Rost, 2004; Strobl, 2012 for detailed information). The requirement that the instrument should measure at ranges of the ability scale where test subjects are located can be verified by item-person maps comparing the ability distribution of respondents with the item difficulties. This is possible because, as already mentioned, ability and item difficulty are measured on the same scale. At model level, unidimensionality or local stochastic independence is verified by comparing restricted with more complex models and judged on the basis of likelihood ratio tests or related statistics such as the Aikaike information criterion (AIC) or the Bayiesian information criterion (BIC) (cf. Burnham & Anderson, 2004). At item level, many diagnostic statistics are based on a comparison of model-estimated probabilities with the empirically observed ones. If there are large differences, mis- or underspecification of the measurement model can be assumed – apart from the latent trait, other factors influence response behavior. In this regard, differential item functioning (DIF) might be virulent. DIF occurs when different groups of respondents (such as men and women, students from different countries etc.) have different probabilities of solving certain test items while controlling for personal ability. DIF is an indicator for a lacking "fairness" of test items (see de Ayala, 2009 for a more detailed discussion). In order to test for DIF, there are different approaches. An intuitive procedure (de Ayala, 2009; Zumbo, 1999) is computing logistic regression models with the items as dependent variables and ability as well as the background characteristics susceptible for DIF as independent variables.

3 A Pilot Study on the Measurement of Competencies in Empirical Social Research Methods

The aim of our study consisted in defining a structural model for competencies in empirical social research methods, in developing test items in order to measure these competencies, in conducting an empirical survey among sociology students, and finally in developing and evaluating the characteristics of the resulting competence scales. While the results concerning the latter point will be presented in chapter 4, in this chapter, we shall address the first three issues.

3.1 Defining Competencies in Empirical Social Research Methods

Structural models of competencies seek to identify and denominate the dimensions of a competence construct (Hartig & Klieme, 2006: 132). Ideally, competence models should be theoretically founded and applicable in empirical measurements. Following Koeppen, Hartig, Klieme, and Leutner (2008), the development of sound theoretical models of competencies is difficult and only very few established models in the whole research on competence assessment do exist. As regards the contents of such models, the definitions are always normative and not a question of "right" or "wrong". Generally, the development of such models can be based on analysis of curricula, module descriptions, content-relevant textbooks, course schedules, interviews with experts, or definitions or agreements of relevant boards, committees, or associations.

Our structural model is organized along two axes, the first one being *cognitive processes* describing what the test person is intended to perform, and the second one being *content-related* topics, objects or situations. Following this conceptualization, the description of competencies always consists of a verb and an object. The cognitive processes of our model follow the conceptualization by Bloom (1956) and Krathwohl (2002) and comprise the levels *knowledge/understanding*, *application/interpretation*, and *evaluation/selection/constructing*. It is assumed that the three processes are organized hierarchically, so for example application or interpretation of a method or a result implies knowledge and understanding of it. This hierarchy, however, must not be confounded with competence levels and is not equal to the notion of difficulty of test items, because the order of the three cognitive processes is conceivable at different competence or difficulty levels.[4] In order to define the second axis, content-related topics, we analyzed a sample of curricula, module descriptions, and course schedules of several German sociology institutes involved in the teaching of empirical social research methods, textbooks

[4] That is, a test item can measure the knowledge of a complex issue and thus be more difficult than another item that asks for the evaluation of a simple problem.

on empirical research methods (for example, Diekmann, 2008; Kühnel & Krebs, 2012; Schnell, Hill, & Esser, 2005), and the official recommendations of the section on methods of empirical social research of the German Sociological Association (Deutsche Gesellschaft für Soziologie – DGS) regarding the formation in research methods. Additionally, we investigated empirical articles in recent volumes of four important German sociology journals[5] with respect to the employment and frequency of data analysis methods in order to separate important and regular used methods from more exotic ones.

A relatively clear-cut result of this investigation is that competencies in empirical social research methods comprise the two main dimensions "data collection methods" and "data analysis methods (including statistics)" (see also Pötschke & Simonson, 2003: 74). Each of the two dimensions includes several sub-dimensions which are depicted in Table 1. The last three columns symbolize that at each combination of content and cognitive process, different difficulty levels are imaginable. For the sub-dimension "statistical laws and properties", there is no cognitive process "evaluate/select/construct", because we consider this task as being the field of duty for mathematicians and not primarily sociologists. As the next step, the cells describing the combination of content, cognitive process and difficulty level were filled with test items. This procedure is described in the next section.

Table 1: Structural Model of Competencies in Empirical Social Research Methods

Dimension	**Sub-Dimension**	**Cognitive Process**	**Difficulty**		
			Easy	**Middle**	**Hard**
Data collection	Survey designs	Know/understand			
		Apply			
		Evaluate/select			
	Sampling	Know/understand			
		Apply			
		Evaluate/select			
	Methods of collecting data	Know/understand			
		Apply			
		Evaluate/select/construct			
	Measurement and scaling	Know/understand			
		Apply/interpret			
		Evaluate/select/construct			
Data analysis & statistics	Statistical laws and properties, notably statistical significance	Know/understand			
		Apply			
	Data analysis procedures and technical issues (software)	Know/understand			
		Apply			
		Evaluate/select/construct			
	Results of empirical analyses (univariate, bivariate, multivariate)	Know/understand			
		Apply/interpret			
		Evaluate/select/construct			

[5] Kölner Zeitschrift für Soziologie und Sozialpsychologie (KZfSS), Zeitschrift für Soziologie (ZfS), Methoden, Daten, Analysen (MDA) and the Berliner Journal für Soziologie (BJS).

Admittedly, the way we defined and presented a structural model for competencies in empirical social research methods is just a rough outline of how one would proceed in a large-scale study; this shortcut is owed to the explorative character of our study and certainly represents a key desideratum for future research. We do think however, that generally the domain of empirical social research methods is well amenable to a consensual definition of a structural model of competencies.

3.2 Test Items and Questionnaire Design

Employing the scheme of Table 1, test items were developed. Ideally, it would have been desirable to cover every cell of the last three columns of Table 1 by several test items. Due to the limited resources of our study, however, this was not possible.[6] Experience from the first pilot study showed that presuming a time frame of 45 minutes for the completion of the whole questionnaire corresponds to a number 30 to 35 test items that are administrable by the test persons. Taking also into account the uncertainty about whether the field phase would be successful and yield enough number of cases, we decided to include 48 items, 24 of which, respectively, relate to data collection and data analysis. The items were administered using a booklet design (Frey, Hartig, & Rupp, 2009) in which six different booklets (questionnaire versions) were randomly administered to the respondents. Each booklet contains 32 test items, divided into four testlets of eight questions. The order of the testlets was permuted, as illustrated in Table 2.

Table 2: Booklet Design of the Survey

Booklet		**1**	**2**	**3**	**4**	**5**	**6**
Testlet	Position 1	A, B	C, D	A, B	C, D	E, F	E, F
	Position 2	C, D	E, F	E, F	A, B	C, D	A, B

Note: A, C, and E are testlets containing eight items each for the dimension data collection. B, D, and F accordingly contain items for the dimension data analysis.

The 48 items split into 22 items with open response format and 26 items with closed response format. Five items of the latter group are assignment or rearrangement items, 17 are single-choice items and four are multiple-choice items. Figure 2 shows three example items (originally in German and translated into English for this article). The first one is intended to measure "evaluation and selection" in the sub-dimension "sampling" on an easy to intermediate level. The second item is situated in the field "application of statistical laws and properties" and its difficulty is considered to be intermediate to hard. The last example item is intended to measure "application of data analysis procedures" on an easy to intermediate level.

[6] If one attempted to cover each cell in the last three columns of Table 1 with, four test items, then the whole number of items had to be $3 \times 20 \times 4 = 240$.

Figure 2: Example Items

FC4 **You want to make a survey of heroin addicts in your hometown and cover preferably all the members in that population. Which sampling method would be ideal?**

- Registration-Based Sampling ☐ 1
- Snowball Sampling ☐ 2
- Cluster Sampling ☐ 3
- Random-Sample ☐ 4

FD2 **Complete the absolute frequencies in the following crosstabulation so that a value of 1 (perfect statistical link) for a Chi-Square based measure of association is reached.**

	no Abitur	Abitur	Row total
No newspaper subscriber			
Newspaper subscriber			
Column total			100

FF2 **Calculate mode, median and arithmetic mean for this group of students.**

Student:	**1**	**2**	**3**	**4**	**5**
Age:	**20**	**20**	**25**	**30**	**35**

a) Mode: ☐

b) Median: ☐

c) Arithmetic mean: ☐

Apart from the test items, the questionnaires also contained questions on other topics. At the beginning, a first set of questions was devoted to variables related to the studies of the students such as course and degree of studies, duration of the studies, or grades. Also, items for the measurement of self-efficacy beliefs (Jerusalem & Schwarzer, 2012) and for a subjective assessment of competencies in empirical social research methods were part of this questionnaire section. The second part contained the test items, followed by a third part on motivational aspects regarding social sciences, and on socio-demographics. Two methodic questions concluded the questionnaire; one item asked the respondent how much effort he or she has – subjectively – invested in filling out the test items, and the last item contained an anonymous, person-specific code (first number of the birthday, first letter of the mother's first name and so on) in order to identify duplicate cases.

3.3 Field Phase and Data Basis

After a pretest, the questionnaires were administered to students in sociological courses at seven universities in Germany and Switzerland. Field phase lasted from October 2013 to January 2014. Because of the explorative character of the study, no efforts were made regarding the sampling; we used a convenience sample of sociology students and mainly profited from colleagues that supported our project and agreed to distribute the questionnaires in their courses. Before distributing the questionnaires, the students were informed about the aims of the study, anonymity, and the voluntary character of taking part in the test. Furthermore (and this was also indicated on the questionnaire), respondents were explicitly instructed to leave items blank for which they did not know the answer. Finally, the students were advised that in case they are still in lower semesters, it could happen that they do not know the correct answer for many items, that this presents no problem, and that they should not be disappointed and continue in answering as many test items as possible.

Because of the undefined population, we are not able to give information regarding the ratio of dropouts. All in all, however, respondents cooperated well. As already our first pilot study had shown, for a good cooperation of the students it is in our view essential not to exceed a time limit of 45 minutes for the completion of the questionnaire and to distribute it at the beginning of the lesson.

The resulting gross sample contains N = 776 cases. For all analysis presented in the subsequent sections, we first excluded duplicate cases. These duplicates arose in case a student filled out the questionnaire in different courses at the same university. The respective cases were identified by the anonymous, personalized code at the end of each questionnaire. Also, the university, gender, and year of birth were taken into account to identify duplicates. Following this procedure, we eliminated 28 observations. In a second step, 11 cases that responded to less than 29 (of 32) test items were excluded, too, but only if the concerning respondents had not responded to the questions that immediately followed the test items (in case they had, we assumed that the test items not responded to are "valid non-response" in the sense that respondents did not know the answer and followed the instructions to leave blank test items for which they did not know the answer). Finally, two observations where the respondent was the teacher or tutor were deleted. Hence, the analysis sample contains 735 observations.

4 A Scale for Measuring Competencies in Empirical Social Research Methods

4.1 Coding of the Test Items and Preliminary Analyses

In a first step, all questionnaires were manually typed into a data file. All responses to the test items were entered as indicated on the questionnaire (including open questions) with the exception of four items that were rated as wrong or correct directly in the course of data entry, because the respondents' answers to these items could not be transferred into a data matrix (for example, items where respondents had to paint a graph or a table). Afterwards, all subtasks of the other test items were rated as correctly answered, wrongly answered, or not answered. In order to assure intercoder reliability, this was done using preliminarily written down solutions for all items in which it was precisely indicated which answers had to be coded as correct. For item analysis, all blank items (not answered) were coded as incorrect answers, because the test instructions explicitly asked all respondents to leave blank items for which the respondent did not know the answer. In the next step, items that encompassed several subtasks were coded into ordinal variables (later analyzed using the partial credit model), summing up all correct responses to the subtasks of the respective items. This procedure is problematic, however, for items where categories of the resulting ordinal variable only have low number of cases or for items where a correct response to $k-1$ subtasks (k being the number of subtasks) perfectly determines the complete score of the item (this concerns matching items, for example, where a correct match of, say, three subtasks implies a correct matching of the fourth subtask). Consequently, we further collapsed the subtasks and coded 12 out of 14 polytomous items into three response categories and two items into four response categories.

Table 3 shows descriptive analyses of the test items. Due to the booklet design, the number of valid answers (N) differs by item groups. The fraction of items that were not answered ranges between 12 percent for item D6 and 89 percent for item B8. A comparison of the average number of blank items by gender (not documented in Table 3) indicates a value of 13 for male students and a value of 15.2 for female students, a difference which is only marginally significant ($t = 1.87$, $p < 0.1$). A regression model that regresses the number of missing values (for all items) on the number of semesters studied in social sciences and the number of courses attended in empirical methods yields an R^2 statistic of 0.4. In our view, this shows that the test instruction to leave blank items for which the answer was not known, worked well. Table 3 also depicts the fraction of correct answers for the (subtasks of the) items. For the sub-dimension “methods of data collection”, the most difficult item is category 2 of the ordinal item C3 which was answered correctly by 7.1 percent of the respondents. The easiest item is item A7 (category 2), answered correctly by 65 percent of the students. The mean of correct answers is 33.5 percent (counting the categories of ordinal items separately). For the sub-dimension “statistics and data analysis”, the most difficult item is item B8, correctly answered by 4.3 percent of the respondents; the easiest items are items D4 and F7 (category 2), answered correctly by 49.6 percent of the students. Here, the mean of correct answers is 24.4 percent, showing that the items for this sub-dimension are more difficult than the one for data collection.

Table 3: Descriptive Analysis of the Test Items

Item	N	Category	% Missing	% Correct	Item	N	Category	% Missing	% Correct
A1	492	1	27.2	24.6	B1	492	1	39.8	44.5
		2		39.6	B2	492	1	44.9	29.1
A2	492	1	52.9	30.7	B3	492	1	61.4	16.9
A3	492	1	66.7	13.4	B4	492	1	76.8	17.9
A4	492	1	23.8	55.7	B5	492	1	73.4	15.2
A5	492	1	37.6	19.1	B6	492	1	51.8	32.1
A6	492	1	36.0	30.9	B7	492	1	55.5	16.7
A7	492	1	15.0	17.5			2		27.9
		2		65.0	B8	492	1	88.6	4.3
A8	492	1	73.4	23.2	D1	492	1	48.4	28.5
C1	492	1	16.7	24.6	D2	492	1	65.7	18.7
		2		53.7	D3	492	1	21.3	38.8
C2	492	1	35.6	34.6	D4	492	1	40.0	49.6
C3	492	1	76.0	13.6	D5	492	1	81.9	4.9
		2		7.1	D6	492	1	12.0	34.2
C4	492	1	18.9	57.9			2		31.1
C5	492	1	32.7	33.5			3		16.1
C6	492	1	30.1	32.9	D7	492	1	56.3	25.6
		2		36.6	D8	492	1	51.4	37.2
C7	492	1	23.8	38.0	F1	486	1	47.9	24.5
		2		36.2			2		19.3
C8	492	1	28.3	57.9	F2	486	1	39.3	36.2
E1	486	1	45.7	35.0			2		16.1
E2	486	1	19.8	27.8	F3	486	1	86.4	6.2
		2		23.5	F4	486	1	77.6	15.8
		3		27.8	F5	486	1	65.4	31.1
E3	486	1	14.8	39.5	F6	486	1	61.5	21.8
		2		32.3	F7	486	1	31.9	18.5
E4	486	1	63.2	28.2			2		49.6
E5	486	1	32.9	27.8	F8	486	1	87.5	4.7
E6	486	1	21.4	64.6					
E7	486	1	62.4	7.4					
		2		22.6					
E8	486	1	23.7	55.4					

Note: Percent missing refers to the fraction of respondents that did not answer the item or respectively at least one of the subtasks of multiple-task-items. Item D3 does not figure in the final scale for data analysis.

4.2 Scale Construction

The scales for competencies in methods of data collection and data analysis were constructed by fitting (generalized) partial credit models in which dichotomous and polytomous items are entered into the model simultaneously.[7] We evaluated the fit of the models by (1) comparing a partial credit specification (constant slopes for every item) with a generalized partial credit specification (separate slope estimated for every item) using likelihood ratio (LR) tests and the BIC statistic, (2) investigating the correlations of the item score with the ability estimate, (3) the outfit and infit statistic (de Ayala, 2009: 51ff.; Linacre, 2002; Wilson, 2005: 127ff.), (4) and the visual comparison between observed and model-estimated ICCs. These steps were reiterated until satisfactory scales were obtained. All IRT analyses were conducted using the gsem procedure in Stata, the TAM procedure for R (Kiefer, Alexander, & Wu, 2014), and the IRTPRO software (Paek & Han, 2013).

Table 4 shows global fit statistics for the two subscales. For both, the GPCM specification shows a significant better fit than the PCM specification. For the scale data collection however, the BIC statistic is in favor of the PCM model, and also visual inspection of the empirical and estimated ICCs and the outfit and infit statistics (see below) argue for choosing the PC specification. In contrast, for the subscale data analysis a PCM showed considerable misfit of some of the items (outfit statistic larger than 1.5, infit statistics larger than 1.3), for which reason the GPCM specification was retained. One item (D3) that exhibited a strong misfit even after modeling a GPCM was excluded. The reliability estimates also depicted in Table 4 have satisfactory values with the exception that the WLE reliability for data collection is only 0.656.

Table 4: Results of the IRT Analysis for the Scales "Data Collection" and "Data Analysis"

	Scale Data Collection	Scale Data Analysis
Initial number of items	24	24
Number of items retained	24	23
−2LL (PCM)	16071,6	11373,6
−2LL (GPCM)	15980,0	11076,5
χ^2 (PCM vs. GPCM) (df)	91,6 *** (23)	297,1 *** (22)
BIC (PCM)	16302,6	11571,6
BIC (GPCM)	16362,8	11419,7
Model chosen	PCM	GPCM
EAP reliability	0,828	0,835
WLE reliability	0,797	0,656
N	735	735

Note: PCM refers to partial credit model and GPCM refers to generalized partial credit model.

[7] For dichotomous items, the (G)PCM reduces to the 1PL or, respectively, the 2PL model.

The item parameters, outfit and infit statistics, and correlations between item and ability estimate are depicted in Tables A1 and A2 in the appendix. Both outfit and infit are measures that account for differences between observed and expected answers to the test items. Outfit and infit range from 0 to infinity with a value of 1 for no discrepancies between empirical and model-based estimates (de Ayala, 2009). Several suggestions do exist regarding which values to accept for outfit and infit. De Ayala (2009: 53) proposes to accept statistics between $1 \pm 2/N^{0.5}$ for infit and $1 \pm 6/N^{0.5}$ for outfit; Wilson (2005: 129) suggests accepting values between 0.75 and 1.33. For the subscale data collection, all fit statistics are in an acceptable range indicating no considerable item misfit. For the subscale "data analysis", some items have exceeding values for some of the items in the outfit statistics. As the infit statistics, however, show no considerable misfit, we decided not to remove the items. Regarding the correlations between item and ability score, Pohl and Carstensen (2012: 11f.) suggest coefficients of greater than 0.3 as good, which applies for all items in our two sub-scales. For some of the polytomous items, one can observe "unordered transition locations" (for example, item A1). This means that the transition point from category score 0 to category score 1 is located at a higher value on the θ scale than the transition point from category score 1 to 2. Although this seems to be counter-intuitive, it does not constitute a problem or misspecification in the (G)PCM (de Ayala, 2009: 166–168).

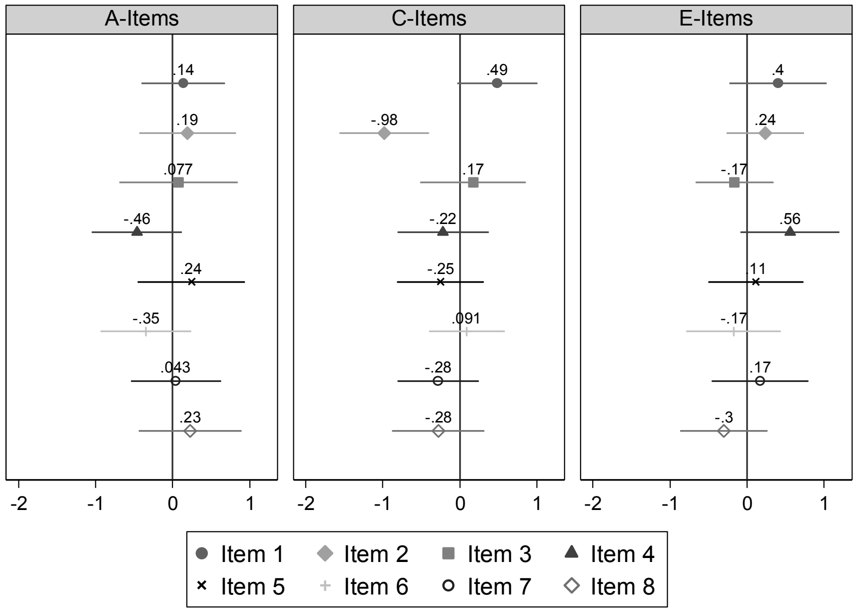
In a final step, we evaluated the scales with respect to differential item functioning (DIF). DIF occurs if the probability of answering an item correctly depends on other respondent characteristics than person ability (Wilson, 2005: 163). We tested the above presented scales for DIF as a function of gender, course of studies (main subject sociology versus other), and university. To do so, we followed the recommendation by Zumbo (1999; see also: de Ayala 2009: 331ff.) and fitted ordinal logistic regression models of the probability of answering each item correctly on the ability estimate, the DIF variable in question, and the interaction between ability estimate and DIF variable. In case the DIF variable executes a significant main effect on the probability of solving the test item while controlling for ability, uniform DIF is indicated for the item under concern. If the interaction effect is significant, non-uniform DIF is virulent, meaning that the relationship between ability and answer probability varies by subgroups. Because also substantially negligible effects tend to become significant with growing sample sizes, Zumbo (1999) proposes to investigate as a measure of effect size the incremental R^2 of the model including the DIF variable as compared to a model without it. Only if the incremental R^2 statistic – Zumbo refers to Nagelkerke's R^2 – is higher than 0.13, a significant DIF effect should be considered as noteworthy.[8]

As concerns DIF as a function of gender, we found no evidence for non-uniform DIF in both subscales. For one item of the subscale data analysis, the interaction coefficient between gender and ability was significant, but neither the incremental R^2 nor the gender-specific difference in predicted probabilities, which we examined using a conditional effects plot, showed a considerable amount of non-uniform DIF. Therefore, we proceeded in examining uniform DIF only. For illustration, figures 3a and 3b present the results of the respective analyses. The figures show the logit coefficients and their 99 percent confidence intervals of the gender main effects. For the subscale data collection, the results

[8] Several authors additionally note the problem of multiple testing or superelevation of the alpha error in the context of multiple significance tests (Kubinger & Draxler, 2007): Even if no significant deviations hold for a given population, taking a five percent alpha level as a basis for assessing significance yields five out of 100 effects significant by chance. Therefore, Zumbo (1999: 27) and others (Kubinger & Draxler, 2007: 138) suggest taking a one percent alpha level.

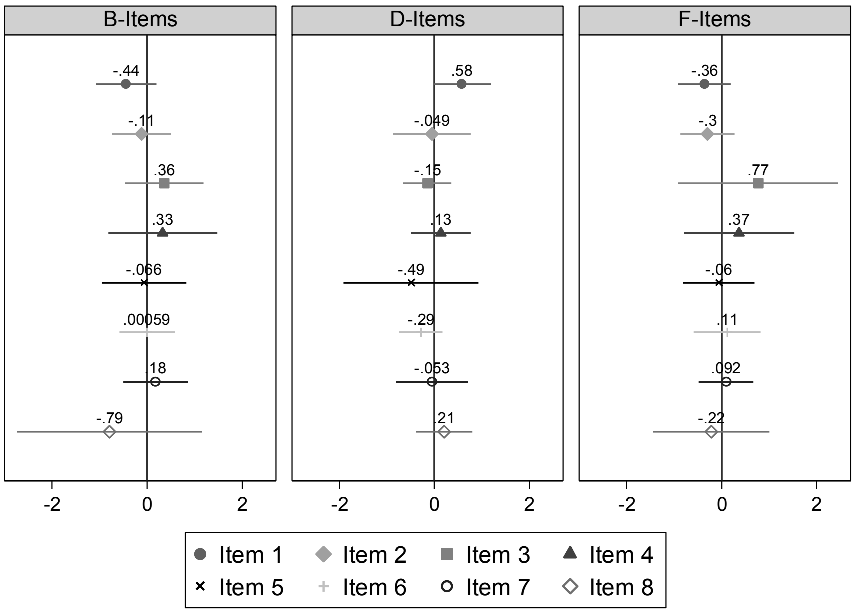
indicate significant DIF for one item (C2) only. However, the effect size of the effect as measured by the incremental R^2 (which takes a value of 0.044) does not indicate substantial DIF, so we did not remove the item from the scale.[9] For the subscale data analysis, no item shows DIF. Taken together, the results suggest that our test is equally fair for both men and women.

Figure 3a: DIF as a Function of Gender (Scale Data Collection)



Notes: The figure shows the magnitude and 99 percent confidence intervals of unstandardized regression coefficients of gender (1 = female) from ordinal logistic regression models regressing the probability of answering each item correctly on the ability estimate and gender.

Figure 3b: DIF as a Function of Gender (Scale Data Analysis)



Notes: see Figure 3. Item D3 does not figure in the final scale for data analysis.

[9] Also, as regards the content of item C2, we found no reason why the item should be subject to gender-specific DIF.

For DIF in dependence of course of studies (students with sociology as main subject versus all others), there is non-uniform DIF for one item in the subscale data collection (not documented). Again, further inspection revealed that the substantial significance is negligible. Uniform DIF occurs for five out of 24 items. As for gender, the incremental R^2 values do not indicate a considerable effect size, so we left the items in the scale. For the subscale data analysis, we found neither non-uniform, nor uniform DIF. As regards DIF in function of the university (here, results are documented in Wolter & Schiener, 2015), we found several items (30 out of 329 tests) with significant main effects of the university (and five out of 329 significance tests yielded significant interaction effects). However, none of them had a large effect size, so we again did not remove items from the scales. Furhermore, there were no items that had systematic strong effects for all universities, and no university showed systematic DIF in one direction (positive or negative) for all items affected. This means that our test does not disadvantage or favor certain universities. Nonetheless, our results here are certainly not as clear-cut as regarding gender specific DIF: In analyses that compare universities, one should bear in mind that the results could be affected by DIF. This could argue against the possibility to employ our scales for the evaluation of higher education institutions, because for this purpose one would ideally ask for tests that function equally in all institutions. We doubt however, that it is possible to succeed in developing tests that fulfill this demand, because the curricula in higher education are highly unstandardized (in contrast to the general education system). On the other hand, one could also argue that for this latter reason, it is not appropriate at all to claim for equally functioning test items because this would thwart the idea of specialization that underlies higher education teaching. Therefore, DIF in dependence of universities could also be judged positively, because it makes visible certain specializations of the respective faculties as regards the contents of their study programs. Altogether, the issue of DIF depending on course of studies and university should be generally discussed in another paper (see our remarks on this topic in Wolter & Schiener, 2015, too).

As a final step of scale construction, the person parameters of the resulting (G)PCM for data collection methods and data analysis were saved using weighted likelihood estimation (Warm, 1989). The ability estimates were then standardized to a mean of 10 and a standard deviation of 5 for subsequent analyses. Furthermore, for some of the analyses presented in the next section, we collapsed the two indicators to a single "global competence indicator" by adding up the two subscales and, again, standardizing the resulting variable to a mean of 10 with standard deviation 5.[10]

In the (G)PCM, item difficulties (or location parameters) and ability are measured on the same scale. This permits to depict both item and the θ distribution in one graph, as implemented in Figures 4a and 4b. For the scale "data collection", the items fit the ability distribution quite well, meaning that the test items are located in areas of θ where most of the students are situated. The items of the subscale "data analysis", by contrast, are too difficult for the θ distribution. Several items measure at high levels of θ where very few students are situated, and there are too few items that measure at low levels of ability.

In large-scale applications of competence assessment, one could investigate the content of the ordered items and formulate levels or steps of ability verbally describing what tasks students at certain θ levels are able to perform. We do not have enough space to attempt this in this paper, and, also, the number

[10] Analyses not documented here show that a two-dimensional model fits the data better than a uni-dimensional one does; however, the two sub-dimensions intercorrelate highly.

of test items is somewhat limited, but, however, at first sight, the difficulty order of our test items makes some sense. Items of the scale data collection, for example, which are located around one standard deviation below the mean measure basic knowledge of elementary concepts such as knowledge of levels of measurement. Items located near the mean measure established knowledge of elementary and more complex concepts and the ability to make decisions regarding sample designs or data coding on an intermediate level. More difficult items located around one standard deviation above the mean measure knowledge of complex and detailed concepts and the ability to develop strategies regarding problems of survey interviews, sampling and others. For the scale data analysis, the abilities measured by items are also interpretable in terms of a growing complexity of tasks. But here, as already mentioned above, our test was too difficult for the sociology students interviewed in our study.

Figure 3a: Item-Person-Map of the Scale Data Collection

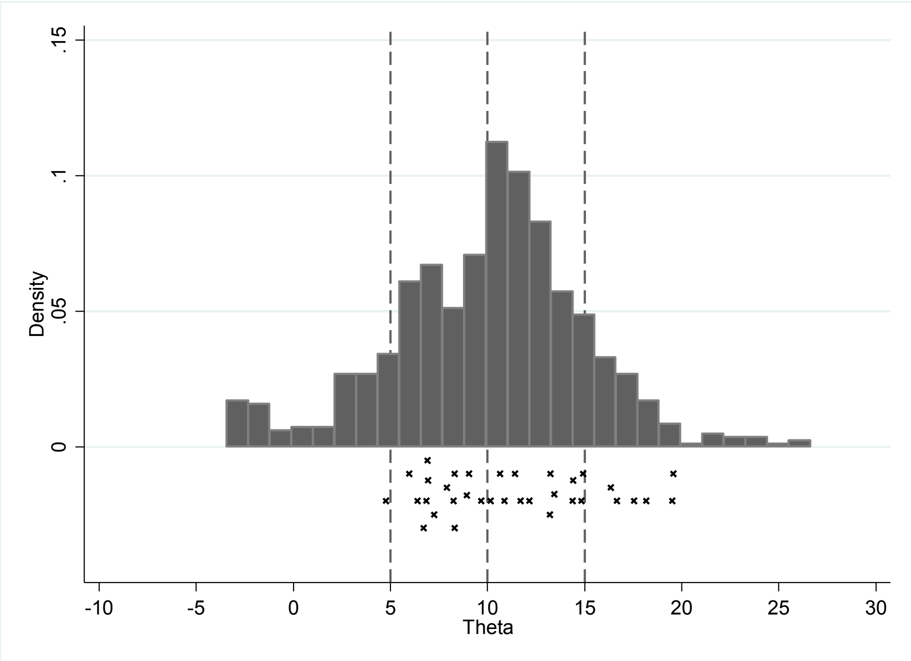
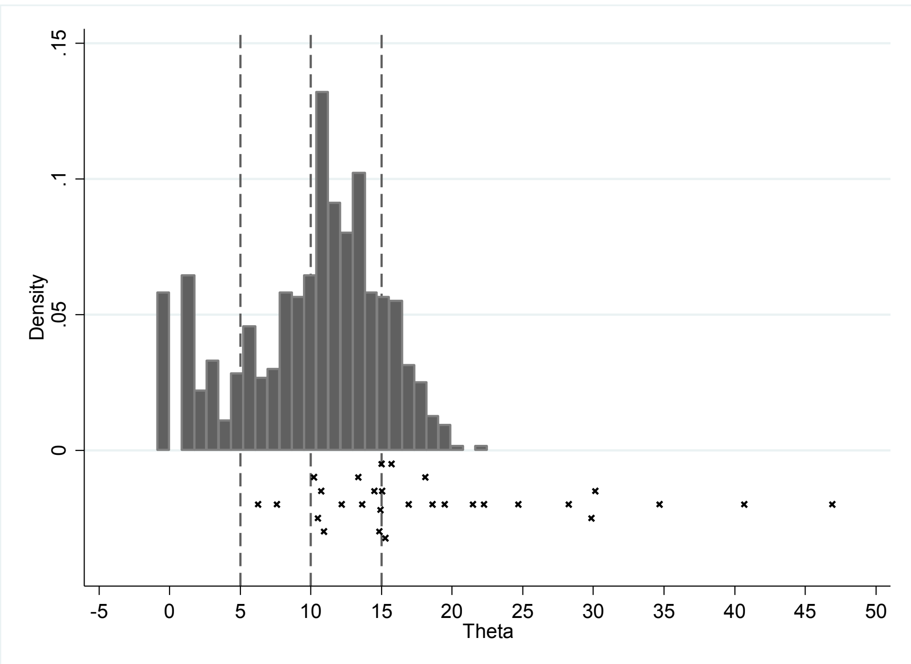


Figure 3b: Item-Person-Map of the Scale Data Analysis



4.3 External Validity and Analysis Potential

In what follows, we will investigate the external validity of the measures presented above and illustrate the analysis potential of the data by presenting content-related results.[11]

Table 5 summarizes descriptive characteristics of the variables entering the subsequent analyses. Two indicators, self-efficacy beliefs and intrinsic motivation measure what Roth (1971) subsumes under the term “self-competence”, which stands for abilities of self-regulation such as learning strategies, motivation or evaluation competencies (Klieme & Hartig, 2007: 20). Self-efficacy beliefs “refer to appraisals of one’s own competencies to plan and execute actions in a successful way in order to achieve desired goals” (Beierlein, Kemper, Kovaleva, & Rammstedt, 2013: 251; see also: Jerusalem & Schwarzer, 2012). The indicator for self-efficacy beliefs employed here is a summed-up index of six

î ▌最鐊ほ髦 ! }ȣ Ĥ 醍Ŧ Ā莖□v梳□v ⌂Ŷ婜

he indicator for intrinsic motivation has been developed by the authors and refers to intrinsic interest in sociological issues or topics (for instance, voluntarily read sociological literature). The variable consists of four unidimensionally loading items that have been summed up. The mean of the grade in the “Abitur” has, on the German scale that reaches from 1 (best) to 5 (worst), a value of 2.4 with standard deviation 0.56. The grades of Swiss students that are also part of the sample have been recoded to the German scale. The variable “parents’ education” has been generated using the general and professional educational degree of the students’ parents, which were approximated in educational years, summed up, and, in case indications for both parents were available, divided by two. 17 percent of the students have a migration background, defined here as being born abroad or having at least one parent that was born abroad. Finally, as a control variable, we asked the respondents at the end of the questionnaire how much effort they had invested in filling out the test items.

In order to avoid that listwise deletion of missing values for multivariate analyses reduces the number of cases too much, we employed multiple imputation to impute missing values. Following the recommendations in StataCorp (2013), missing values were imputed into 20 new imputed data sets using chained equations.

[11] Also see Wolter and Schiener (2015), where we present additional findings from multilevel models, reporting effects of the universities on competence achievement.

Table 5: Variables of the Empirical Analyses

Variable	Notes	M	STD	N
Ability “data collection methods”	See section 4.2	10	5	735
Ability “data analysis methods”	See section 4.2	10	5	735
Global ability	See section 4.2, sum of the subscales for “data collection” and “data analysis”	10	5	735
Semesters in social sciences	Semesters attended in social sciences [0…19]	4.92	2.84	728
Number of attended courses in empirical research methods	0 to 4 or more courses	2.08	1.26	679
Course of studies	0 = main subject sociology, 1 = other subject	0.33		729
Self-efficacy beliefs	1 = low to 7 = high (see text)	4.31	1.05	701
Intrinsic motivation	1 = low to 7 = high (see text)	4.46	1.19	691
Exam grade in the Abitur	1 = very good to 5 = insufficient (German grade system)	2.41	0.56	718
Employment besides studying	1 = continuously/often, 0 = never/occasionally	0.53		706
Parents’ education	In years (see text)	14.57	2.56	674
Migration background	1 = yes, 0 = no (see text)	0.17		688
Gender	1 = female, 0 = male	0.62		703
Effort in answering the test items	Subjectively rated effort in answering the test items; 1 = little effort to 7 = very high effort	4.13	1.55	703

Note: M = mean, STD = standard deviation, N = number of valid cases.

Let us first examine the external validity of our competence indicators. If our ability estimates really measure what they are intended to, then they should correlate high with the advancement of the studies of the respondents. Regression models for the two subscales in Table 6 show the effects of the duration of studies (number of semesters in social sciences) and the number of attended courses in research methods, on the ability estimate. All reported effects are highly significant, and the R^2 statistics show that a good part of the variance of the competence indicators are explained by the variable on studies advance: The number of attended semesters in social sciences and the number of courses in research methods explain 29 percent of the variance for data collection and 35 percent for data analysis. We interpret these results in favor of the external validity of our indicators: They do measure what students – at different universities – learn during their studies.

As can be seen from the effects of the squared term of the semester variable, its effect is curvilinear: Competencies in research methods tend to grow faster during the first semesters of studies than in later ones; furthermore, there is a maximum turning point at about 10–12 semesters, after that, the competencies tend to decline again. Although this seems to be counterintuitive, it is not implausible for several reasons: Due to forgetting, advanced students could be worse than younger ones that have just passed the relevant courses on research methods. Also, there is certainly selectivity in the sample, because students were tested *during actual courses* and students that are still present in courses after six years of studies are presumably worse than those who have already finished their studies by that time.

Table 6: Determinants of Competencies in Empirical Social Research Methods

	Data Collection	Data Analysis	Global Competence
Semesters in social sciences (centered)	0.407 *** (0.080)	0.268 ** (0.078)	0.486 *** (0.074)
Semesters in social sciences (squared from centered)	−0.059 *** (0.012)	−0.026 * (0.012)	−0.051 *** (0.011)
Number of courses in research methods	1.667 *** (0.154)	2.079 *** (0.151)	1.253 *** (0.160)
Course of studies (1 = minor subject)			−2.516 *** (0.373)
Self-efficacy beliefs			0.378 ** (0.139)
Intrinsic motivation			0.086 (0.124)
Abitur grade			−1.465 *** (0.269)
Employment (1 = yes)			0.034 (0.282)
Parents' education			−0.050 (0.054)
Migration background (1 = yes)			−0.632 + (0.368)
Gender (1 = female)			−0.982 ** (0.304)
Effort in answering test items			0.526 *** (0.092)
Constant	7.061 *** (0.385)	5.951 *** (0.374)	9.443 *** (1.590)
R^2 (corr.)	0.292	0.352	0.482
N	735	735	735

Note: Linear regression using multiple imputation of missing values. Unstandardized regression coefficients and their robust standard errors in brackets. +: p < 0,1; *: p < 0,05; **: p < 0,01; ***: p < 0,001.

For the last model M3 in Table 6 we use the global competence indicator as dependent variable (the sum of the two subscales). The results indicate that, as one expects, that students whose main subject is not sociology are half a standard deviation worse than students with main subject sociology. Regarding our measures for self-efficacy beliefs and intrinsic motivation, only the former exerts a significant positive effect on competencies. More remarkable, however, is the finding that the exam grade in the Abitur significantly affects, other things being equal, the success at university: One point on the German grade system from five to one yields more than a quarter standard deviation improvement in our competence measure. Contrarily, the fact whether students are pursuing an employment in addition to their studies, does not affect competence. One could have supposed a negative effect here, because jobbing students have supposedly less time for their studies than those who are not working. Also, the parental background measured here by the education of the students' parent does not influence achievement at university. Having a migration background tends to have a negative, though only marginally significant effect. As a matter of fact, however, female students are nearly a fifth standard deviation worse than male students – a result that we already found and speculated about in our first pilot study (Wolter & Schiener, 2014). This result is also in line with findings from other studies on competence assessment (Förster, Happ, & Zlatkin-Troitschanskaia, 2012; Walstad & Robson, 1997). Although this issue should be more widely discussed in another paper, we see prima facie three mechanisms that could explain this result: First, the effect could be substantial, meaning that female students *have* lower competencies than men. Secondly, as Förster et al. (2012) and Spiel, Schober, and Litzenberger (2008) suppose, women tend to show a higher risk-aversion than men, resulting in the fact that when facing test items for which they are only partly sure about the correct answer, women tend more often than men to leave the item blank while men try to guess in those cases. Third, the results could also be explained by gender-specific DIF, that is, test items that are more difficult for female students under control of the ability they are supposed to measure. For our study, we found no DIF as a function of gender, and the number of not answered items is only slightly higher (on a 10 percent alpha level) for women, which points to the "substantial effect-hypothesis." One could argue then that large scale assessments have repeatedly found that female students have a lower ability than male ones as concerns mathematical tasks, while being better in language-related tasks (e.g., Mullis et al., 2012; Prenzel, Sälzer, Klieme, & Köller, 2013). As our test that we presented in this article focuses to some extents on statistical/mathematical issues, our findings could be in line with this interpretation.

Finally, the last effect in model M3 (Table 6) is the expectable positive effect of the subjectively rated effort in answering the test items. We used this variable as a control variable in order to at least partly avoid that the other effects in the model are biased by differing effort in answering the test item in dependence of the other variables in the model. As a matter of fact, however, the effects of the other variables turned out to have almost exactly the same effects when we dropped the effort variable from the model.

5 Discussion

The aim of the article was to present results and insights from an empirical pilot study devoted to the development of a measuring instrument assessing competencies in quantitative empirical social research methods. The motivation for such a project stems from a general demand for objective measures of learning achievement which in turn results from several flaws of general evaluation practices applied today – mostly subjective measures or input instead of output measures. While this demand has already led to the development and establishment of measurement instruments for pupils in the general education sector (cf. PISA et al.), this is not the case for the higher education sector. Meanwhile, however, the methodology for such projects is well-established and empirical studies show their general feasibility, so there seems to be no reason why one should not attempt to transfer the methodology from the general to the higher education sector.

Altogether, our results demonstrate the general feasibility of developing satisfactory scales for the measurement of competencies in quantitative empirical social research methods. Analyses on the relationship between the competence indicators and several predictor variables demonstrate a good external validity of the scale and a potential for fruitful content-related analyses.

Of course, our indicators are far from being ready for substantial field use and represent only a first step of developing more elaborated scales. More concretely, there are several open issues that should be addressed in future research. First, we did not explore the potential of multidimensional IRT models for assessing the dimensionality of the construct. Second, as already mentioned, the issue of DIF as a function of course of study and university should be generally discussed and empirically evaluated with regard to the potential of using competence measures like the one presented here for evaluation of higher education institutions. Third, and related to the last point, one should also discuss whether it makes sense to measure competencies of differently advanced students with one test. The PISA approach, for instance, is explicitly defined for a narrowly defined age group of pupils and based on official definitions of curricula. On the contrary, the PIAAC survey for the assessment of adults' competencies also uses an approach comparable to the one presented in this paper and measures different age groups with one test. In this regard, one should, fourth, also attempt to conduct longitudinal measurements of competencies. A fifth desideratum is to further develop a verbal competence-level-model defining levels of competence and the tasks that students with certain levels are able to carry out.

A further area of problems are sampling difficulties and causality concerns. These, however, are not specific to competence measurements as understood in this article, but rather pertain to all kinds of measurement of student achievement. Sampling problems such as selective dropouts and/or lacking cooperation of institutions, the teaching staff or students can bias results if dropouts are correlated to the variables under concern (Wolbring, 2013). Causality concerns arise if one attempts to isolate the *causal* effects of institutions, curricula, political interventions, and so on – which represents the first and foremost goal of any evaluation analysis. A huge literature on this issue exists in the field of research into school-effectiveness and the "value added" of educational institutions (see, for instance, Rutter & Maughan, 2002; Tekwe et al., 2004; Timmermans, Snijders, & Bosker, 2013). A simple comparison of, for example, higher education institutions, might be misleading, because the true causal effect is biased by the fact that students change institutions, study paths, have different prior abilities (different levels of competence before entering the respective university), etc. Possible solutions for these issues are widely discussed in the literature, and are challenging – minimal requirements seem to be longitudinal data, a measurement of prior ability, and multilevel analysis.

To finish: The undertaking of defining and measuring competencies in higher education is, alone because of the normative character of structural competence models, necessarily in a tenuous position. On the other hand, what is the alternative? Given the numerous flaws of other evaluation measures and practices adhered to today, there is a clear demand for objective competence measures. Therefore, we clearly plead for further pursuing research into these issues, which should first and foremost be guided by pragmatism. Our results show that with reasonable effort, one is able to develop competence scales that can be used as a basis for an objective measurement of student achievement. Also, we think that this especially holds for the domain of empirical social research methods which is well suited for being made accessible to an objective measurement of competencies.

Acknowledgments

This work was funded by the Internal University Research Funding and the Center for Educational and Higher Educational Research (ZBH) of the University of Mainz. Part of the research was carried out within the framework of a project seminar at the University of Mainz. We thank our students who participated in the project seminar and our colleagues that supported us with the data collection.

Appendix

Table A1: Item Parameters for the Scale "Methods of Data Collection"

Item	Cat.	Location	(SE)	Slope	Outfit	(t)	Infit	(t)	r
A1	1	0.037	0.113	1	0.808	−1.429	0.905	−1.891	0,676
	2	−0.085	0.110	1	0.812	−1.066	0.913	−1.969	
A2		1.049	0.107	1	0.934	−0.710	0.955	−0.949	0,500
A3		2.279	0.141	1	0.869	−0.704	0.974	−0.266	0,347
A4		−0.023	0.101	1	0.977	−0.349	0.984	−0.388	0,503
A5		1.794	0.124	1	1.093	0.677	1.014	0.228	0,390
A6		1.037	0.107	1	1.156	1.685	1.002	0.061	0,434
A7	1	−0.734	0.143	1	0.781	−0.531	0.981	−0.210	0,587
	2	−1.253	0.113	1	1.207	1.957	1.153	2.646	
A8		1.508	0.116	1	1.005	0.081	0.993	−0.103	0,438
C1	1	−0.756	0.130	1	1.063	0.335	1.040	0.592	0,536
	2	−0.661	0.106	1	1.127	1.256	1.163	3.430	
C2		0.763	0.104	1	0.951	−0.622	1.003	0.073	0,441
C3	1	1.944	0.125	1	0.895	−0.703	0.979	−0.283	0,508
	2	1.586	0.196	1	0.630	−0.129	1.003	0.069	
C4		−0.409	0.101	1	1.004	0.083	1.005	0.140	0,494
C5		0.817	0.105	1	0.942	−0.735	1.008	0.189	0,411
C6	1	−0.501	0.115	1	1.145	0.760	0.983	−0.283	0,560
	2	0.202	0.108	1	1.283	1.236	1.228	4.632	
C7	1	−0.860	0.119	1	0.774	−0.937	0.919	−1.370	0,685
	2	0.333	0.107	1	0.749	−2.043	0.873	−3.051	
C8		−0.409	0.101	1	0.916	−1.393	0.925	−1.936	0,527
E1		0.766	0.105	1	0.952	−0.588	0.920	−1.877	0,520
E2	1	−0.965	0.133	1	0.749	0.565	0.863	−2.018	0,694
	2	0.152	0.113	1	0.932	−0.511	1.051	1.039	
	3	0.400	0.120	1	1.172	0.466	0.937	−1.134	
E3	1	−0.788	0.120	1	1.258	1.497	1.058	1.005	0,521
	2	0.513	0.114	1	1.186	0.896	1.164	3.113	
E4		1.146	0.110	1	0.883	−1.265	0.962	−0.756	0,443

E5		1.170	0.111	1	1.029	0.324	1.014	0.276	0,426
E6		−0.738	0.105	1	0.960	−0.545	0.961	−0.866	0,504
E7	1	2.266	0.120	1	1.085	0.602	1.111	1.766	0,526
	2	−0.420	0.132	1	1.195	0.613	1.127	1.828	
E8		−0.260	0.101	1	1.060	0.963	1.019	0.489	0,458

Note: The column "r" shows the Pearson correlation coefficient between item and latent trait.

Table A2: Item Parameters for the Scale "Statistics & Data Analysis"

Item	Cat.	Location	(SE)	slope	(SE)	Outfit	(t)	Infit	(t)	r
B1	1	0.398	(0.113)	1.817	(0.166)	0.991	(−0.017)	0.987	(−0.235)	0.608
B2	1	1.239	(0.114)	1.429	(0.137)	0.878	(−1.036)	0.996	(−0.063)	0.479
B3	1	2.797	(0.149)	2.356	(0.155)	3.618	(2.077)	0.973	(−0.288)	0.475
B4	1	4.603	(0.191)	4.693	(0.204)	0.805	(33.821)	0.909	(−0.741)	0.559
B5	1	3.372	(0.163)	2.820	(0.162)	0.904	(0.756)	0.974	(−0.256)	0.482
B6	1	0.952	(0.107)	1.139	(0.127)	1.027	(0.335)	1.007	(0.160)	0.438
B7	1	1.039	(0.137)	2.447	(0.142)	1.268	(0.591)	1.053	(0.716)	0.725
	2	2.180	(0.148)	2.447		1.700	(6.508)	1.010	(0.160)	
B8	1	8.651	(0.322)	5.051	(0.210)	0.274	(119.475)	1.315	(1.489)	0.339
D1	1	1.079	(0.108)	0.983	(0.120)	1.030	(0.398)	0.991	(−0.166)	0.394
D2	1	2.610	(0.147)	2.397	(0.157)	0.873	(0.037)	1.004	(0.072)	0.514
D4	1	0.000	(0.111)	1.684	(0.157)	1.348	(2.871)	1.040	(0.793)	0.618
D5	1	5.570	(0.257)	3.152	(0.182)	0.500	(3.370)	1.094	(0.624)	0.335
D6	1	−1.004	(0.126)	0.732	(0.060)	0.881	(−0.690)	0.962	(−0.552)	0.595
	2	−0.898	(0.102)	0.732		1.034	(0.552)	1.019	(0.496)	
	3	0.155	(0.132)	0.732		0.930	(−0.234)	0.993	(−0.063)	
D7	1	1.928	(0.135)	2.384	(0.168)	0.752	(−0.244)	1.017	(0.272)	0.573
D8	1	0.679	(0.107)	1.326	(0.133)	0.943	(−0.526)	0.980	(−0.412)	0.518
F1	1	0.751	(0.111)	1.274	(0.092)	1.112	(1.045)	1.072	(1.362)	0.613
	2	1.821	(0.136)	1.274		0.707	(−0.439)	0.966	(−0.475)	
F2	1	0.054	(0.118)	1.816	(0.111)	0.898	(−0.568)	0.877	(−2.308)	0.695
	2	2.193	(0.151)	1.816		1.096	(0.736)	1.236	(2.714)	
F3	1	7.170	(0.272)	4.522	(0.194)	0.446	(26.860)	1.175	(0.956)	0.391
F4	1	4.663	(0.194)	4.360	(0.193)	0.653	(9.345)	0.956	(−0.346)	0.548
F5	1	1.529	(0.131)	2.485	(0.185)	0.707	(−0.448)	1.013	(0.212)	0.612
F6	1	1.811	(0.127)	1.592	(0.141)	1.019	(0.161)	1.043	(0.641)	0.462
F7	1	−0.062	(0.124)	1.250	(0.099)	0.611	(−1.995)	0.806	(−3.136)	0.723
	2	−0.753	(0.114)	1.250		1.122	(0.774)	1.103	(1.894)	
F8	1	4.219	(0.232)	1.877	(0.178)	1.041	(0.354)	1.057	(0.380)	0.267

Note: The column "r" shows the Pearson correlation coefficient between item and latent trait.

Literature

Baumert, Jürgen, Bos, Wilfried, & Watermann, Rainer. (1998). *TIMSS/III. Schülerleistungen in Mathematik und den Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich. Zusammenfassung deskriptiver Ergebnisse*. Berlin: Max-Planck-Institut für Bildungsforschung.

Becker, Gary S. (1975). *Human Capital. A Theoretical and Empirical Analysis, with Special Reference to Education* (2 ed.). Chicago/London: The University of Chicago Press.

Beierlein, Constanze, Kemper, Christoph J., Kovaleva, Anastassiya, & Rammstedt, Beatrice. (2013). Kurzskala zur Erfassung allgemeiner Selbstwirksamkeitserwartungen (ASKU). *Methoden, Daten, Analysen, 7*(2), 251–278.

Blömeke, Sigrid, Zlatkin-Troitschanskaia, Olga, Kuhn, Christiane, & Fege, Judith. (2013). Modeling and Measuring Competencies in Higher Eduation: Tasks and Challenges. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and Measuring Competencies in Higher Education: Tasks and Challenges* (pp. 1–10). Rotterdam: Sense Publishers.

Bloom, Benjamin S. (Ed.). (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. New York: David McKay.

Burnham, Kenneth P., & Anderson, David A. (2004). Multimodel Inference. Understanding AIC and BIC in Model Selection. *Sociological Methods and Research, 33*(2), 261–304.

Clayson, Dennis E. (2009). Student Evaluations of Teaching: Are They Related to What Students Learn? A Meta-Analysis and Review of the Literature. *Journal of Marketing Education, 31*(1), 16–30.

de Ayala, Rafael J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

Diekmann, Andreas. (2008). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen* (19 ed.). Reinbek bei Hamburg: Rowohlt.

Förster, Manuel, Happ, Roland, & Zlatkin-Troitschanskaia, Olga. (2012). Valide Erfassung des volkswirtschaftlichen Fachwissens von Studierenden der Wirtschaftswissenschaften und der Wirtschaftspädagogik - eine Untersuchung der diagnostischen Eignung des Wirtschaftskundlichen Bildungstests (WBT). *bwp@, Berufs- und Wirtschaftspädagogik – online, 22*, 1–21.

Frey, Andreas, Hartig, Johannes, & Rupp, André A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Hartig, Johannes, & Klieme, Eckhard. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 127–143). Heidelberg: Springer.

Jerusalem, Matthias, & Schwarzer, Ralf. (2012). Dimensionen der Selbstwirksamkeit. In A. Glöckner-Rist (Ed.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 15.00*. Bonn: GESIS.

Kiefer, Thomas, Alexander, Robitzsch., & Wu, margaret. (2014). *Package "TAM", Version 1.0-1*.

Klieme, Eckhard, & Hartig, Johannes. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Eds.), *Kompetenzdiagnostik* (pp. 11–29). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, Eckhard, & Leutner, Detlev. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik, 52*(6), 876–903.

Koeppen, Karoline, Hartig, Johannes, Klieme, Eckhard, & Leutner, Detlev. (2008). Current Issues in Competence Modeling and Assessment. *Zeitschrift für Psychologie / Journal of Psychology, 216*(2), 61–73.

Krathwohl, David R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice, 41*(4), 212–218.

Kubinger, Klaus D., & Draxler, Clemens. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell. *Diagnostica, 53*(3), 131–143.

Kühnel, Steffen-M., & Krebs, Dagmar. (2012). *Statistik für die Sozialwissenschaften: Grundlagen, Methoden, Anwendungen* (6 ed.). Reinbek bei Hamburg: Rowohlt.

Linacre, John M. (2002). What Do Infit and Outfit, Mean-Square and Standardized Mean? *Rasch Measurement Transactions, 16*(2), 878.

Masters, Geoff N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika, 47*(2), 149–174.

Meulemann, Heiner. (2002). Die empirische Sozialforschung in der soziologischen Ausbildung und der Berufspraxis von Soziologen. In U. Engel (Ed.), *Praxisrelevanz der Methodenausbildung (Sozialwissenschaftliche Tagungsberichte Band 5)* (pp. 45–67). Bonn: Informationszentrum Sozialwissenschaften.

Mincer, Jacob. (1974). *Schooling, Experience and Earnings*. New York/London: Columbia University Press (National Bureau of Economic Research).

Moosbrugger, Helfried. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (2 ed., pp. 227–274). Berlin/Heidelberg: Springer.

Müller-Benedict, Volker, & Tsarouha, Elena. (2011). Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen. *Zeitschrift für Soziologie, 40*(5), 388–409.

Mullis, Ina V. S., Martin, Michael O., Foy, Pierre, & Arora, Alka. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill / Amsterdam: TIMSS & PIRLS International Study Center, Lynch School of Education / International Association for the Evaluation of Educational Achievement (IEA).

Muraki, Eiji. (1992). A generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

OECD (Ed.). (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*: OECD Publishing: http://dx.doi.org/10.1787/9789264201118-en.

Organisation for Economic Cooperation and Development (OECD) (Ed.). (2013). *Assessment of Higher Education Learning Outcomes (AHELO). Feasibility Study Report Volume 2 – Data Analysis and National Experiences*: OECD.

Paek, Insu, & Han, Kyung T. (2013). IRTPRO 2.1 for Windows (Item Response Theory for Patient-Reported Outcomes). *Applied Psychological Measurement, 37*(3), 242–252.

Pant, Hans Anand, Böhme, Katrin, & Köller, Olaf. (2012). Das Kompetenzkonzept der Bildungsstandards und die Entwicklung von Kompetenzstufenmodellen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (pp. 49–55). Münster: Waxmann.

Pohl, Steffi, & Carstensen, Claus H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests*. Bamberg: University of Bamberg.

Pötschke, Manuela, & Simonson, Julia. (2003). Konträr und ungenügend? Ansprüche an Inhalt und Qualität einer sozialwissenschaftlichen Methodenausbildung. *ZA-Informationen / Zentralarchiv für Empirische Sozialforschung, 52*, 77–92.

Prenzel, Manfred, Artelt, Cordula, Baumert, Jürgen, Blum, Werner, Hammann, Marcus, Klieme, Eckhard, . . . (PISA-Konsortium Deutschland) (Eds.). (2008). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich*. Münster: Waxmann.

Prenzel, Manfred, Sälzer, Christine, Klieme, Eckhard, & Köller, Olaf. (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Rasch, Georg. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Kopenhagen: The Danish Institute for Educational Research.

Rost, Jürgen. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2 ed.). Bern: Hans Huber.

Roth, Heinrich. (1971). *Pädagogische Anthropologie. Band 2: Entwicklung und Erziehung. Grundlagen einer Entwicklungspädagogik*. Hannover: Schroedel.

Rutter, Michael, & Maughan, Barbara. (2002). School Effectiveness Findings 1979–2002. *Journal of School Psychology, 40*(6), 451–475.

Samejima, Fumiko. (1969). Estimation of Latent Ability Using A Response Pattern of Graded Scores. *Psychometrika Monograph Supplement, 34*.

Schnell, Rainer. (2002). Ausmaß und Ursachen des Mangels an quantitativ qualifizierten Absolventen sozialwissenschaftlicher Studiengänge. In U. Engel (Ed.), *Praxisrelevanz der Methodenausbildung (Sozialwissenschaftliche Tagungsberichte Band 5)* (pp. 35–44). Bonn: Informationszentrum Sozialwissenschaften.

Schnell, Rainer, Hill, Paul B., & Esser, Elke. (2005). *Methoden der empirischen Sozialforschung* (7 ed.). München: Oldenbourg.

Spiel, Christiane, Schober, Barbara, & Litzenberger, Margarete. (2008). *Projektbericht: Evaluation der Eignungstests für das Medizinstudium in Österreich*. Wien: Universität Wien (Evaluationsprojekt im Auftrag des Bundesministeriums für Wissenschaft und Forschung).

StataCorp (Ed.). (2013). *Stata Multiple-Imputation Reference Manual Release 13*. College Station, Texas: StataCorp LP.

Stehle, Sebastian, Spinath, Birgit, & Kadmon, Martina. (2012). Measuring Teaching Effectiveness: Correspondence Between Students' Evaluations of Teaching and Different Measures of Student Learning. *Research in Higher Education, 53*(8), 888–904.

Strobl, Carolin. (2012). *Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis* (2 ed.). München, Mehring: Rainer Hampp Verlag.

Tekwe, Carmen D., Carter, Randy L., Ma, Chang-Xing, Algina, James, Lucas, Maurice E., Roth, Jeffrey, . . . Resnick, Michael B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–36.

Timmermans, Anneke C., Snijders, Tom A. B., & Bosker, Roel J. (2013). In Search of Value Added in the Case of Complex School Effects. *Educational and Psychological Measurement, 73*(2), 210–228.

Tremblay, Karine. (2013). OECD Asessment of Higher Education Learning Outcomes (AHELO): Rationale, Challenges and Initial Insights from the Feasability Study. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and Measuring Competencies in Higher Education: Tasks and Challenges* (pp. 113–126). Rotterdam: Sense Publishers.

Tremblay, Karine, Lalancette, Diane, & Roseveare, Deborah. (2012). *Assessment of Higher Education Learning Outcomes (AHELO). Feasibility Study Report Volume 1 – Design and Implementation*: OECD.

Walstad, William B., & Robson, Denise. (1997). Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. *The Journal of Economic Education, 28*(2), 155–171.

Warm, Thomas A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54*(3), 427–450.

Wilson, Mark. (2005). *Constructiong Measures. An Item Response Modeling Approach*. Mahwah: Erlbaum.

Wolbring, Tobias. (2013). *Fallstricke der Lehrevaluation. Möglichkeiten und Grenzen der Messbarkeit von Lehrqualität*. Frankfurt/New York: Campus.

Wolter, Felix, & Schiener, Jürgen. (2014). Auf dem Weg zum „Hochschul-PISA“? Zur Messung „soziologischer Kompetenzen“. *Soziale Welt, 65*(1), 47–73.

Wolter, Felix, & Schiener, Jürgen. (2015). Kompetenzdiagnostik an der Hochschule. Methoden, Herausforderungen und aktuelle Entwicklungen. In D. Großmann & T. Wolbring (Eds.), *Evaluation von Studium und Lehre. Grundlagen, methodische Herausforderungen und Lösungsansätze*. Wiesbaden: Springer VS (to be published in 2015).

Zlatkin-Troitschanskaia, Olga, Blömeke, Sigrid, Kuhn, Christiane, & Buchholtz, Christiane. (2012). Wirksamkeitsprüfungen im Hochschulbereich – Aufgaben und Herausforderungen des BMBF-Forschungsprogramms "Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor". *Zeitschrift für Evaluation, 11*(1), 95–103.

Zumbo, Bruno D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.