Gess, Christopher; Wessels, Insa; Blömeke, Sigrid

Domain-specificity of research competencies in the social sciences: Evidence from differential item functioning

Journal for educational research online 9 (2017) 2, S. 11-36

Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 9 (2017), No. 2, 11–36

Christopher Gess, Insa Wessels & Sigrid Blömeke

Domain-specificity of research competencies in the social sciences: Evidence from differential item functioning

Abstract

To investigate the domain-specificity of research competencies, higher education students from the social sciences were assessed with a standardized test in four disciplines: (a) sociology, (b) political science, (c) educational studies, and (d) psychology. The measure covered declarative and procedural knowledge of research methods, methodology, and procedures. Quantitative and qualitative research traditions were represented equally by test items. The domain-specificity of the measure was examined by detecting and explaining differential item functioning (DIF) between the disciplines. It was hypothesized that due to differences in opportunities to learn (OTL), students from different disciplines responded differently to subgroups of items. As expected based on the OTL-patterns, research traditions significantly explained variance in DIF. While psychology students were more likely to correctly answer items addressing quantitative methods than students with the same overall ability level but from different disciplines, students of all other disciplines were more likely to solve items addressing qualitative methods. These differences coincided with differences in OTL. Overall, the findings suggest that research competencies are similar across the social sciences, but differences between disciplines exist in their focus on quantitative or qualitative methods.

Keywords

Higher education; Achievement test; Research competence; Domain-specificity; Differential item functioning

Christopher Gess, MPA (corresponding author) · Insa Wessels, M.Sc., bologna.lab, Humboldt-Universität zu Berlin, Hausvogteiplatz 5-7, 10099 Berlin, Germany
e-mail: christopher.gess@hu-berlin.de
insa.wessels@hu-berlin.de

Prof. Dr. Sigrid Blömeke, Centre for Educational Measurement (CEMO), Faculty of Education, University of Oslo, Niels Henrik Abels hus, 0318 Oslo, Norway
e-mail: sigribl@cemo.uio.no

Domänenspezifität sozialwissenschaftlicher Forschungskompetenz: Analyse von Differential Item Functioning

Zusammenfassung

Um die Domänenspezifität von Forschungskompetenz innerhalb der Sozialwissenschaften zu untersuchen, wurde ein Test zur Messung von Forschungskompetenz in Sozial-, Politik-, Bildungswissenschaften und Psychologie eingesetzt. In diesem Kompetenztest wurde deklaratives und prozedurales Wissen zu quantitativen und qualitativen Forschungsmethoden sowie übergreifendes Forschungsprozesswissen erhoben. Testaufgaben zu quantitativen und qualitativen Forschungsmethoden waren zu gleicher Anzahl vorhanden. Differential Item Functioning (DIF) zwischen den Studienfächern wurde analysiert. Es wurde erwartet, dass sich Unterschiede in den Lerngelegenheiten (OTL) zwischen den Fächern im Antwortverhalten der Studierenden widerspiegeln. Wie erwartet konnte auf Basis von Unterschieden in OTL ein substantieller Anteil an Varianz zwischen DIF-Parametern der Testaufgaben erklärt werden. Psychologiestudierende zeigten eine relative Stärke in quantitativen Methoden, die Studierenden der anderen Studienfächer bei qualitativen Methoden. Die beobachteten Stärken spiegeln erwartungsgemäß die curricularen Schwerpunkte und OTL der Studienfächer wider. Zusammenfassend weisen die Ergebnisse darauf hin, dass die Bestandteile von Forschungskompetenz zwar weitgehend fachübergreifend sind, sich jedoch Unterschiede zwischen den Studienfächern zeigen, je nachdem, ob der Schwerpunkt auf quantitative oder qualitative Forschungsmethoden gelegt wird.

Schlagworte

Kompetenzmessung; Hochschuldidaktik; Forschungskompetenz; Domänenspezifität; Differential Item Functioning

1. Introduction

Acquiring research competencies (RC) is an important goal of higher education (British Academy, 2012; Wissenschaftsrat, 2006). Yet, in the social sciences, there is a lack of discussion about specific learning objectives, and no comprehensive tools to assess the attainment of these objectives are available (Earley, 2014).

Existing competency measures focus on parts of the research process only (information literacy, see Katz, 2007; statistical literacy, see Stone, 2006) or on understanding and applying research results (Groß Ophoff, Schladitz, Leuders, Leuders, & Wirtz, 2015). Additionally, none of the measures available incorporates the assessment of competencies in both quantitative and qualitative social research. While in the natural sciences, recent efforts yielded a measure applicable across disciplines (in physics, chemistry and biology, see Hartmann, Upmeier zu

Belzen, Krüger, & Pant, 2015), no such measure is available across the social sciences.

Against this backdrop, for the present paper, a newly developed measure (Gess, Geiger, & Ziegler, in press) assessing RC across social research paradigms and social-scientific disciplines was applied to examine the domain-specificity of RC. Students' RC is typically taught in discipline-specific learning environments, such as specific research methods courses for students enrolled in psychology or educational studies (Wagner, Garner, & Kawulich, 2011). The opportunities to learn (OTL) may vary between disciplines, presumably resulting in distinct response behaviors to the measure's items.

The study follows, thus, a recommendation initially made for research on scientific reasoning to focus on *differences* between disciplines instead of their *commonalities* (Fischer et al., 2014). Hence, differences in OTL between the disciplines were analyzed and related to between-discipline variance found in the measure's items. This analysis allows for a closer investigation of the new measure but also provides tentative conclusions on the nature of social-scientific RC as such.

2. Conceptual framework and state of research

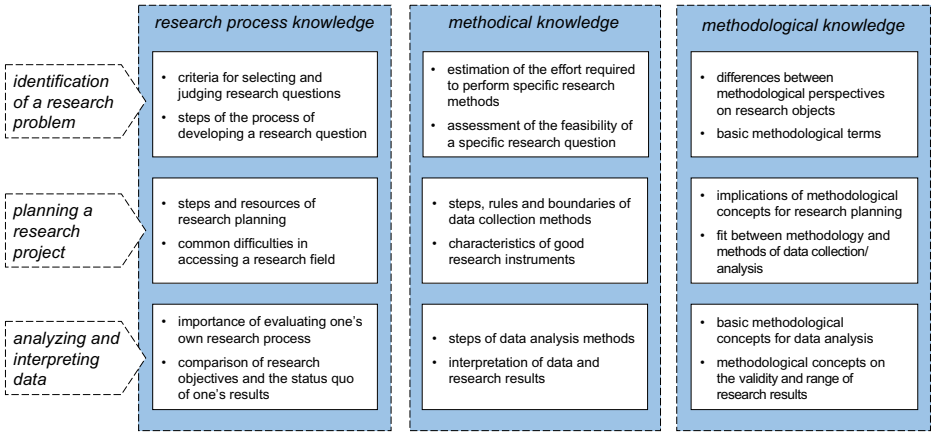
2.1 Research competencies in social sciences

The term *research competencies* is used in different ways. First of all, it is important to differentiate between studies that focus on an engagement *in* research and on an engagement *with* research (Borg, 2010). The competencies to engage with research are often examined in higher education programs that train for a specific profession (e.g. teacher, nursing or medical education). In these studies, students are tested on their competencies to understand and use research results in professional decision making ("educational research literacy", see Groß Ophoff et al., 2015, p. 560). In the present paper, the competencies to engage *in* research – i.e. the competencies necessary to generate new knowledge based on scientific methods – are examined. In the following, this line of research is called *research competency*.

Our study built on a conceptual framework of research competency. The framework specified the content areas associated with RC in the social sciences – based on interviews and two surveys with methodological experts. Experts highlighted three *research steps* as particularly insightful: (a) finding and defining a research problem, (b) planning a research project and (c) analyzing and interpreting data. Three *knowledge domains* were identified as "cognitive dispositions" (Koeppen, Hartig, Klieme, & Leutner, 2008, p. 68) underlying successful social research: (a) research process knowledge, (b) knowledge of research methods and (c) knowledge of methodologies. By highlighting knowledge domains and research steps, the competency model is in line with the continuum of approaches to model competencies

(Blömeke, Gustafsson, & Shavelson, 2015). For each combination of knowledge domain and research step, indicators were identified (see Figure 1). Note that content-specific knowledge pertaining to the research field is required to competently conduct a research project. However, this is not included as the model is intended to be valid across research topics and disciplines.

Figure 1: Model of social-scientific research competency. Rows represent different research steps, columns represent different knowledge domains, and the content of the cells represents the respective indicators.



2.2 Measuring research competencies in social sciences

The model of RC in social sciences was transformed into a test instrument that was intended to be used to evaluate opportunities to learn social research, i.e. courses on research methods and gain insights on the nature of RC. Its target population are both undergraduate and graduate students majoring in a social science. In order to capture the breadth of the underlying construct, for each combination of knowledge domain and research step (see Figure 1) three items were included in the final measure. The items addressed different cognitive processes: recalling knowledge, applying knowledge, evaluating research projects (Anderson & Krathwohl, 2001). In total, 3x3x3 = 27 items were included in the final measure (sample items can be found in Annex A). Items on both quantitative and qualitative research methods were included in the measure. While items addressing the first knowledge domain (research process knowledge) were contingent to both research traditions, items addressing knowledge of research methods and knowledge of methodologies either focused on quantitative or qualitative research methods alternately. In total, the measure included nine quantitative and nine qualitative items. The item development was based on textbook material as well as on an expert survey (n = 8).

To ensure content validity, the items have been refined in an item panel (Wilson, 2005) and in a think-aloud study with students from social sciences (n = 14). Finally, the items were validated in an expert rating with experts in research and teaching in sociology, political science, educational studies and psychology (n = 24), as well as with knowledge constructing professionals (working with evaluation, market research, opinion and social research, consulting, n = 56). In this expert rating, the model of RC proved relevant both from the perspective of university experts and professionals and the quality of all items included in the final measure was judged acceptable on average.

2.3 Measuring research-related opportunities to learn in social sciences

In order to develop a measure that assesses learning opportunities to acquire skills to perform social-scientific research, relevant content domains were identified using syllabi from research methods courses and textbook material. These sources were chosen in order to assess the OTL as close to actual teaching content as possible. In a second step, these content domains were grouped and classified according to the knowledge domains embodied in the model of RC: (a) research process knowledge, (b) knowledge of research methods and (c) knowledge of methodologies. Additionally, (for the latter two knowledge domains) the content domains were differentiated between qualitative and quantitative knowledge. However, deviating from the model of RC, the content domains could not be assigned to single research steps as in syllabi and textbooks they were often introduced across research steps.

Having identified, grouped and classified these content domains, items were constructed in a forth step: Following the prompt "Regarding the following list of topics, please mark each topic with a cross in case you have ever studied it, i.e. it was taught in a course", students rated a list of 27 content domains as yes or no (see the items and descriptive statistics in Annex B). In total, quantitative and qualitative methods and methodologies were represented by 11 content domains each (see Table 1).

Table 1: Content domains used in the construction of the research-related OTL measure

	research process knowledge	knowledge on research methods	knowledge on methodologies	total
quantitative research	–	8	3	11
qualitative research	–	5	6	11
contingent to both traditions	5	–	–	5
total	5	13	9	27

2.4 Differences in research education between social-scientific disciplines

Research on the teaching of RC in different social-scientific disciplines is rare (Earley, 2014). Discipline-specific emphases in research education have especially not been examined. Yet, specifications and guidelines for curricula allow some tentative conclusions. The American Psychological Association [APA] (2011) suggests a research education that builds on statistical and experimental methods and does not refer to qualitative methods – which is mirrored in psychology curricula (Peden & Carroll, 2009). The American Sociological Association on the contrary advises to engage students in both quantitative and qualitative research and consequently defined learning goals for both research traditions (McKinney, Howery, Strand, Kain, & Berheide, 2004). The American Political Science Association [APSA] (2004) recommends exposure to qualitative and quantitative, descriptive, interpretive and explanatory approaches in research. While for educational studies, no comparable guidelines seem to be available, the German Educational Research Association (DGFE, 2004) also advocates for implementing both quantitative and qualitative methods in curricula.

These guidelines for curricula show a major difference between psychology and other social-scientific disciplines. While in sociology, political science and educational studies the guidelines comprised both quantitative and qualitative methods, qualitative approaches are neglected in psychology. Although these conclusions are based on US-American guidelines, they also seem to apply to the German context. The emphasis of quantitative approaches in German psychology education is reflected in methodological textbooks (Hussy, Schreier, & Echterhoff, 2010; Mruck & Mey, 2000). The psychological research education focuses predominantly on statistics and a hypothetico-deductive methodological framework (Lettau & Breuer, 2007). However, qualitative methods are important in some areas of psychological research (e.g. in educational psychology, see Butler, 2006; in psychotherapy research, see Lutz & Knox, 2014; in test construction, see Ziegler, Kemper, & Lenzner, 2015). In line with this, psychological experts that reviewed the test instrument presented in this paper also attached importance to aspects and items referring to qualitative methods and methodologies.

However, although the RC model can be regarded domain-general given these evaluations, disciplines emphasize different aspects of RC, which should result in discipline-specific strengths and weaknesses on the corresponding items. To detect these, one can use a measurement tool known as differential item functioning (DIF). DIF occurs if groups of tested persons – e.g., psychologists versus sociologists – systematically perform higher or lower on an item than expected given their mean performance on the total score (Shepard, 1982). This conception is similar to international competency assessments, where broad constructs are used that are generally homogenous so that it is not possible or meaningful to report subscores for subsets of items but country-specific achievement profiles evoked by underlying

multidimensionality in small parts of the constructs are analyzed to retrieve the information about strengths and weaknesses (e.g., Klieme & Baumert, 2001).

3. Research questions and hypotheses

The main objective of this paper was to analyze any potential between-discipline variance in the performance on certain items while controlling for the students' overall performance on the measure. This variance was assessed in relation to differences in content domains highlighted – as reflected in OTL – in four disciplines: (a) sociology, (b) political science, (c) educational studies and (d) psychology. We expected that students of a specific discipline systematically outperform other students in items that address content domains highlighted in the respective discipline. This overall expectation translates into five specific hypotheses:

Given the distinctive methodical and methodological emphases evident in the disciplines' guidelines for curricula (see Section 2.4) our first hypothesis was that between-discipline DIF in items exists (H1). Secondly, we expected that students of sociology, political science and educational studies report a higher share of OTL in qualitative methods and methodologies compared to psychology students (H2) and hence a lower share in quantitative methods and methodologies. Thirdly, if these two hypotheses hold true, we expected that students of sociology, political science and educational studies show strengths and weaknesses on the same items when compared to psychology students (H3). More specifically, students of these disciplines should outperform psychology students on items addressing qualitative research methods (H4) and underperform on items addressing quantitative methods (H5) even if their overall RC score is controlled for.

4. Methodology

4.1 Sample

The sample was drawn in clusters as the test was administered in regular university classes. Selection of classes for assessment was based on the following criteria: Four groups of classes were sampled according to their target population (Bachelor students in their 3rd and 5th semester and Master students in their 1st and 3rd semester). For each discipline and for each of these groups, at least two classes were randomly selected. More than two classes were drawn in case of expected small class sizes. Lecturers were approached via email. If access to a class was denied, a replacement was randomly drawn. In total, 103 lecturers were approached via email and 51 classes taught by 50 different lecturers were assessed in five universities in four out of the 16 states of Germany. Each discipline was assessed in at least three universities.

The sample of eligible participants comprised 681 students. Six students stated not to be proficient enough in German and were excluded. The remaining sample comprised 675 students (see Table 2). The number of students in the sample differed between disciplines due to differing sizes of classes and attendance in class (*min-max* = 1-71, *M* = 13.2, *SD* = 11.8) and proportions of non-eligible participants (e.g. students studying a social science as a minor subject). Survey weights were used to accommodate for differences in sample sizes between the disciplines and thus to prevent overweighting a discipline. Within the sample of students, a subsample was randomly drawn that – after taking the test – filled out an additional survey on the students' OTL (*n* = 309).

The collection of data was administered in class by three trained supervisors. Completing the RC-test took around 35 minutes, the verbal and mostly standardized introduction took around two minutes and the supplementary survey questions took around eight minutes.

Table 2: Disciplines and year of study of the sample (*N* = 675)

	bachelor (undergraduates)			master (graduates)			diploma/ magister	total
	2nd year	3rd year	Adv.	1st year	2nd year	Adv.		
sociology	33	24	12	27	14	12	1	123
political science	63	31	14	43	28	8	1	188
educational studies	26	23	2	61	19	4	3	138
psychology	74	38	4	76	28	5	1	226
total	196	116	32	207	89	29	6	675

Note. 3 students pursuing a Master's degree did not report the semester. They are reported in column "Adv." (3+ years in BA/BSc or 2+ years in MA/MSc programmes).

4.2 Dimensionality of the measures and validation of test score interpretations

4.2.1 Research competencies

The dimensionality of the measure was assessed by comparing four multidimensional models to the unidimensional model. The multidimensional models were grounded in the test construction guidelines (see Section 2.2). RC-1 differentiated between the knowledge domains addressed (methodical, methodological and research process knowledge). RC-2 differentiated between the research traditions addressed (quantitative methods, qualitative methods, research process knowledge contingent to both traditions). RC-3 differentiated between the three research steps addressed (problem identification, research planning, data analysis and interpretation). RC-4 differentiated between cognitive processes addressed (recall-

Table 3: Comparison of the unidimensional and four multidimensional models of RC (n = 675)

model	*AIC*	*BIC*	*Log Lik.*	$\chi^2(3)$
RC-0 (unidimensional)	20373.03	20616.82	-10132.51	–
RC-1 (knowledge domains)	20379.01	20636.35	-10132.50	0.02^{ns}
RC-2 (research traditions)	20376.06	20633.39	-10131.03	2.98^{ns}
RC-3 (research steps)	20376.60	20633.94	-10131.30	2.43^{ns}
RC-4 (cognitive processes)	20376.59	20633.93	-10131.30	2.44^{ns}

Note. ^{ns} = not significant

ing knowledge, applying knowledge, critiquing research projects). A 2PL between-item multidimensional IRT-scaling was performed (Hartig & Höhler, 2009) using the R-package mirt (Chalmers, 2012). Models RC-1 to RC-4 were compared to the unidimensional model RC-0 using a likelihood ratio test (LRT). Additionally, the information criteria AIC and BIC were interpreted. None of the multidimensional models showed better model fit than the unidimensional model according to AIC and BIC (see Table 3). Among the multidimensional models, the best fit was reached by the model differentiating between research traditions (RC-2). However, RC-2 did not show significantly better model fit than RC-0 according to the LR-Test. Similarly, all other multidimensional models did not show significantly better model fit than RC-0 (see Table 3). As the unidimensional model demonstrated acceptable fit ($\chi^2(324) = 342.11$, $p = .23$, $RMSEA = .009$ $CI_{90\%}[< .001, .017]$, $CFI = .948$) in a CFA using the WLSMV-estimator for categorical indicators and taking the clustered data structure into account (type = complex, Muthén & Muthén, 2015), a unidimensional interpretation of the measure seems adequate.

In this unidimensional model, all 27 items showed good fit ($0.95 \leq wMNSQ \leq 1.07$) according to Adams and Wu (2002). The reliability based on the weighted likelihood estimation (WLE; Warm, 1989) was acceptable ($REL_{WLE} = .74$). Score estimates were based on WLE. Students around the mean RC score of 0 (±0.1 *SD*) solved on average 17 of 27 items correctly. Students who scored a standard deviation below the mean (±0.1 *SD*) solved 13 items and students who scored a standard deviation above the mean (±0.1 *SD*) solved 21 items correctly, on average.

According to the intended uses and target population of the measure (see Section 2.2), several validation propositions have been addressed in a separate paper (Gess et al., in press): Evidence suggested that the test instrument measured a learnable, social-scientific construct relevant to research performance: Using cross-sectional data from students of sociology, political science, educational studies and psychology (n = 669), it has been demonstrated that higher RC test scores were related to study progress: Graduates showed significantly higher test scores than undergraduates. Based on a subset of this data that was limited to graduate students (n = 290), it has been demonstrated that RC scores were related to grades in their final BA-theses with incremental validity over and above self-rated research self-

efficacy. In a laboratory study (n = 82), psychology students showed higher test scores than chemistry students even after controlling for verbal-deductive reasoning and general knowledge.

4.2.2 Opportunities to learn

The dimensionality of research-related OTL was assessed in the same way as with RC. Two multidimensional models were compared to a unidimensional model. The multidimensional models were based on the item construction process outlined in Section 2.3. The model OTL-1 differentiated between the knowledge domains addressed in the OTL (methodical, methodological and research process knowledge). OTL-2 differentiated between the research traditions addressed (quantitative methods, qualitative methods, research process knowledge contingent to both traditions). To compare the models OTL-1 and OTL-2 to the unidimensional model OTL-0, the same procedure was used as in Section 4.2.1. Both multidimensional models demonstrated significantly better fit than the unidimensional models, according to the LR-Test (see Table 4). Judging by AIC and BIC, the model differentiating between research traditions (OTL-2) achieved the best fit. In a CFA, this model achieved good fit ($\chi^2(321) = 389.37$, $p = .005$, $RMSEA = .026$ $CI_{90\%}$[.015, .035], $CFI = .955$) and thus a three-dimensional interpretation seems adequate. The reliability was acceptable in all three dimensions (quantitative OTL: $\alpha = .79$, qualitative OTL: $\alpha = .85$, research process: $\alpha = .72$).

Table 4: Comparison of the unidimensional and two multidimensional models of research-related OTL (n = 309)

model	*AIC*	*BIC*	*Log Lik.*	$\chi^2(3)$
OTL-0 (unidimensional)	7706.74	7908.34	-3799.37	–
OTL-1 (knowledge domains)	7497.40	7710.20	-3691.70	215.33***
OTL-2 (research traditions)	7083.74	7296.54	-3484.87	628.99***

Note. OTL were assessed in a subgroup of the sample with n = 328 (see section 4.1). 19 students of this subsample did not answer this additional survey.

*** $p < .001$

4.3 Data analyses

4.3.1 Detection of differences between disciplines in answering items correctly (H1)

Analysis of differential item functioning (DIF) was applied to detect discipline-specific strengths and weaknesses on single items and thus to test hypothesis H1. For this purpose, a reduced DIF-free set of items (so called anchor items) was identified to be used as common metric in DIF detection as high rates of DIF-items in

the metric can lead to Type I errors (Stark, Chernyshenko, & Drasgow, 2006). The set of anchor items was identified using the "forward-MTT method" (Kopf, Zeileis, & Strobl, 2015a, p. 38), which causes the lowest ratio of Type I and II errors among all major anchor selection strategies when iteratively selecting anchor items (ibid.; Kopf, Zeileis, & Strobl, 2015b). It was performed pairwise between all four disciplines using the R-package psychotools (Zeileis, Strobl, Wickelmaier, Komboz, & Kopf, 2016). The final anchor set was based on the items most frequently selected in pairwise anchor sets. In subsequent DIF-analyses, the common metric was based on these anchor items: Ability scores were derived from a 2PL-IRT model using the R-package mirt (Chalmers, 2012). Item-fit and model-fit were evaluated using wMNSQ-values (Adams & Wu, 2002) and the M_2-statistic (Maydeu-Olivares & Joe, 2005).

DIF would be present when students with equal ability levels but from different disciplines differ in their probability of answering an item correctly. For these items, the students' response behavior cannot be fully explained by their average ability but additionally by specific strengths or weaknesses, for example due to differences in their disciplinary background. Two types of DIF could in this case be differentiated: Uniform DIF would be the main effect of the disciplines, whereas nonuniform DIF would be the interaction of disciplines and ability score (e.g., Zumbo, 2007). If an item shows uniform DIF only, students of a discipline systematically outperform other students throughout the ability levels. If an item shows nonuniform DIF, the impact of the discipline on the response behavior is different across ability levels, e.g. only high performing students of a discipline outperform high performing students of other disciplines but no difference is found between low performing students.

DIF was detected using manifest logistic regressions (Swaminathan & Rogers, 1990) which allow to identify both uniform and nonuniform DIF (Hambleton, Swaminathan, & Rogers, 1991). Per item, three logistic regressions were performed. In the first regression, the probability of giving a correct answer was estimated based on the 2PL-IRT test scores of the DIF-free anchor-item set. In the second regression, in addition to the DIF-free test scores, the four disciplines were included as three dummy variables with psychology as the reference group. In the third regression, in addition to the ability score from the DIF-free items and the disciplines, interaction terms of the ability score and disciplines were included. Comparing the first and second model and the second and third model separately as suggested by Güler and Penfield (2009) allowed for detecting the type of DIF (uniform DIF in the first comparison, nonuniform DIF in the second comparison). In model comparison, a χ^2-test with Holm-adjusted p-values (Holm, 1979) was used for assessing significance and the differences of R^2_{adj} as effect sizes.

Negative DIF parameters indicated that the students of that discipline did worse on an item compared to the students in the reference group (psychology students) with the same test score (i.e., the same average ability). Two sets of thresholds have been proposed to classify the magnitude of DIF (Jodoin & Gierl, 2001; Zumbo & Thomas, 1997 as cited in Zumbo, 1999). In the present paper, the more

conservative thresholds by Jodoin and Gierl (2001) were applied ($\Delta R^2 \leq .035$ is negligible, $.035 < \Delta R^2 \leq .07$ is moderate, $.07 < \Delta R^2$ is large DIF).

4.3.2 Identification of differences in opportunities to learn (H2)

Our overall expectation was that students systematically outperform other students in items that address research methods highlighted in the respective discipline. In order to investigate these methodological emphases in the disciplines and hence to test H2, the shares of quantitative and qualitative OTL in all methodical or methodological OTL reported were analyzed. The shares of quantitative OTL were computed by $OTL_{QN,share} = OTL_{QN}/(OTL_{QN}+OTL_{QL})$. The *share* of OTL was used instead of the *mean* amount of OTL as it reveals curricular emphases: If students of discipline A report twice the absolute number of methodical and methodological OTL but the same high share of quantitative OTL as students of discipline B, both disciplines equally emphasize quantitative methods. In this case, students of discipline A should generally outperform students of discipline B but show the same strengths and weaknesses when comparing both to discipline C that highlights qualitative methods.

The shares of quantitative OTL were compared between psychology students and all other students in a *t*-test. To ensure that the OTL shares reported by psychology students differ from those of *each* of the other disciplines, additional pairwise *t*-tests were performed for each combination of disciplines. Holm-correction was used to prevent inflation of Type I errors. *T*-tests on $OTL_{QL,share}$ were obsolete as this share is the converse share of $OTL_{QN,share}$. To estimate the effects sizes, Cohen's *d* was used. Cluster-robust standard errors were applied in all *t*-tests using the R-package rms (Harrell, 2015).

4.3.3 Identification of discipline-specific strengths and weaknesses (H3-H5)

In order to test the assumption that sociology, political science and educational studies showed the same DIF-patterns (H3), the directions of the coefficients were examined and the correlations between them were estimated. The stronger the correlation, the more similar the DIF patterns.

In order to test the hypotheses that students of sociology, political science and educational studies outperform psychology students in items addressing qualitative methods (H4) and underperform in items addressing quantitative methods (H5), two complementary approaches were applied:

The first approach was to *analyze the DIF-parameters by item characteristics*. For each discipline, the DIF-parameters were regressed on two independent variables differentiating qualitative (H4) and quantitative (H5) items. Items contingent to both quantitative and qualitative research traditions were used as baseline in the

regressions. A positive regression coefficient on an item characteristic indicates a strength of the students of the discipline in the respective group of items. All items were used instead of the significant DIF-items only. The latter would have resulted in a small number of units of analysis (c.f. Blömeke, Suhl, & Döhrmann, 2013; Klieme & Baumert, 2001). Thus, a regression with 27 units of analysis was performed separately for sociology, political science and educational studies. Robust regressions that weighted for outliers were used (implemented in the R-package MASS; see Venables & Ripley, 2002). The advantage of this approach is that it is performed within the unidimensional framework that was used in DIF detection. However, it is an indirect assessment of the hypotheses as the DIF-parameters are used as units of analysis.

The second approach took a different perspective by *analyzing the students' response behavior* in separate regression analyses for qualitative (H4) and quantitative (H5) items. The mean performance of students on these groups of items were regressed on the disciplines, controlling for the students' overall DIF-free RC scores. Cluster-robust standard errors were applied using the R-package rms (Harrell, 2015). The advantage of this approach is that it directly linked performance on groups of items to the disciplines. However, as the response behavior was analyzed regarding groups of items, a multidimensional interpretation of RC was applied although a unidimensional interpretation of RC was indicated (see Section 4.2.1). As both approaches have comparative advantages and disadvantages, both were used for reasons of triangulation.

5. Results

5.1 Differences between disciplines in answering items correctly (H1)

DIF-free anchor-item set. The anchor sets identified for pairwise discipline comparisons contained on average 9.8 items (*min* = 7, *max* = 13). The final anchor-item set contained 10 items. All 10 items demonstrated good fit to the scale created through 2PL-IRT modelling (0.96 ≤ *wMNSQ* ≤ 1.05). The model fit was excellent ($\chi^2(35) = 31.98$, $p = .612$, *RMSEA* = .002, *CFI* = 0.998, *SRMSR* = .031; see Maydeu-Olivares, 2013). The reliability was low though, due to the low number of anchor items ($REL_{WLE} = .58$). WLE-estimated ability scores ranged from -6.18 to +3.35 with a mean of 0 and a standard deviation of 1.44 and correlated strongly with the 2PL-IRT ability scores generated using all 27 items ($r = .80$, $p < .001$). Due to missing values on all ten anchor items, three cases had to be excluded from the subsequent analyses. Students with a score around the mean of 0 (±0.1 *SD*) solved seven of ten items correctly, on average. Students scoring around 1 standard deviation (±0.1 *SD*) below the average solved four items, students scoring around 1 standard deviation (±0.1 *SD*) above the average solved nine items on average.

Table 5: Uniform DIF by item (n = 675)

Item	b_{SO}	b_{PO}	b_{ED}	$\Delta\chi^2(3)$ [a]	ΔR^2_{adj} [b]	class.[c]
BM_RP_1	-0.236	0.199	-0.287	4.8	.008	A
BM_RP_2	-0.380	-0.484	-0.606	5.5	.009	A
BM_RP_3	-0.207	0.048	0.259	2.2	.004	A
BM_RP_4	0.389	0.059	0.454	5.7	.011	A
BM_RP_5	0.208	0.162	0.151	0.8	.002	A
BM_RP_6	-0.037	-0.156	0.093	0.6	.001	A
BM_RP_7	0.291	0.015	0.355	2.7	.006	A
BM_RP_8	-0.311	0.389	-0.499	15.3	.031	A
BM_RP_9	0.127	-0.381	-0.258	4.1	.009	A
QN_MD_1	1.066***	0.526*	0.122	23.6**	.046	B
QN_MD_2	-0.751**	-0.775***	-0.753**	18.5*	.036	B
QN_MD_3	-0.769**	-0.689**	-0.546*	12.3	.027	A
QN_MD_4	0.311	0.685*	0.688*	7.1	.008	A
QN_ML_1	-0.290	-0.355	-0.532	3.4	.008	A
QN_ML_2	-0.811**	-0.545*	-0.534*	13.1	.028	A
QN_ML_3	-1.443***	-1.410***	-1.511***	64.5***	.135	C
QN_ML_4	-0.309	-0.408	-0.511*	5.1	.009	A
QN_ML_5	0.310	0.039	-0.256	3.2	.006	A
QL_MD_1	-0.805**	-0.469*	-0.607*	12.6	.027	A
QL_MD_2	0.700**	0.652**	0.419	12.6	.026	A
QL_MD_3	1.207***	0.702**	1.051**	17.2*	.049	B
QL_MD_4	1.326***	1.092***	0.963**	25.2**	.059	B
QL_MD_5	0.569*	0.094	0.886***	18.3*	.037	B
QL_ML_1	-0.406	-0.558*	-0.100	7.3	.016	A
QL_ML_2	0.936**	0.640**	0.497	13.4	.029	A
QL_ML_3	0.232	-0.245	0.209	3.9	.007	A
QL_ML_4	1.020***	0.420	0.633*	15.2	.034	A

Note. bold = anchor item. QN = quantitative methods. QL = qualitative methods. BM = both research traditions. RP = research process knowledge. MD = methodical knowledge. ML = methodological knowledge. b_{SO}, b_{PO}, and b_{ED} are regression coefficients with psychology students as the reference group. SO = sociology. PO = political science. ED = educational studies.

[a] $\Delta\chi^2 = -2(LL_{2nd\ model} - LL_{1st\ model})$. [b] $\Delta R^2_{adj} = R^2_{adj,2nd\ model} - R^2_{adj,1st\ model}$. [c] class. = DIF-classification, A = negligible DIF ($\Delta R^2 \leq .035$), B = moderate DIF ($.035 < \Delta R^2 \leq .07$), C = large DIF ($.07 < \Delta R^2$, Jodoin & Gierl, 2001).

* Holm corrected $p < .05$, ** Holm corrected $p < .01$, *** Holm corrected $p < .001$.

Detection of DIF. The data revealed that 21 out of 27 items did not show significant DIF (see Table 5), including all anchor items. Thus, as expected, DIF was present in some items. According to the classification proposed by Jodoin and Gierl (2001), five items showed moderate and one item strong DIF. In these items, some of the disciplines thus showed a substantially higher or lower probability of answering an item correctly than predicted by the students' ability level. The amount of variance explained by the students' disciplines was small on average (M_{R^2adj} = .025, SD_{R^2adj} = .027). The items showed neither significant nor relevant nonuniform DIF and consequently, the average amount of nonuniform DIF was negligible (M_{R^2adj} = .007, SD_{R^2adj} = .005). Thus, the impact of the ability level on the response behavior did not differ between disciplines (negligible interaction effect of ability and discipline on the probability of solving an item).

5.2 Differences in opportunities to learn (H2)

As expected, the share of OTL (see Table 6) regarding quantitative methods and methodology reported by psychology students (M = 0.82, SD = 0.13) was significantly higher ($t(299)$ = 6.55, p < .001) than that of the combined three other social-scientific disciplines (M = 0.60, SD = 0.17). The effect size of this difference was large (d = 1.38).

Table 6: Mean amount and ratio of opportunities to learn depending on the discipline and research tradition (n = 309)

		sociology (n = 51)	political sc. (n = 85)	educational st. (n = 62)	psychology (n = 111)
		M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Amount of OTL	quantitative (0-11 OTL)	8.63 (2.03)	6.92 (2.75)	8.28 (2.01)	9.76 (2.14)
	qualitative (0-11 OTL)	6.27 (2.89)	4.84 (3.11)	6.51 (2.97)	2.55 (1.22)
	research process (0-5 OTL)	3.76 (1.42)	4.07 (1.19)	3.76 (1.51)	3.26 (1.60)
Relative amount of OTL[a]	share of quantitative OTL in all methodological OTL	.60 (.14)	.62 (.20)	.58 (.15)	.82 (.13)
	share of qualitative OTL in all methodological OTL	.40 (.14)	.38 (.20)	.42 (.15)	.18 (.13)

Note. OTL were assessed in a subgroup of the sample with n = 328 (see section 4.1). 19 students of this subsample did not answer this additional survey.

[a] The share of OTL is calculated on an individual level. Because of missing values, it slightly deviates from the share of the groups' mean amounts of OTL.

This difference was confirmed in pairwise comparisons, the effect sizes were large in all comparisons (see Table 7). Between the other disciplines – sociology, political science and educational studies – no significant differences were found and effect sizes were small to medium. These findings supported the hypothesis H2 that students of sociology, political science and educational studies report a lower share of

quantitative OTL than psychology students and conversely, a higher share of qualitative OTL.

Table 7: Pairwise mean comparisons of the share of quantitative opportunities to learn between the disciplines (n = 309)

	sociology			political science			educational studies		
	df	*t*	*d*	*df*	*t*	*d*	*df*	*t*	*d*
political science	128	-0.87	-0.13						
educational studies	109	0.65	0.13	139	1.60	0.24			
psychology	158	-5.88 ***	-1.66	188	-5.46 ***	-1.20	169	-6.85***	-1.77

Note. Negative *t*- and *d*-values indicate a lower share of OTL regarding quantitative methods and methodology in the column-group compared to the row-group and vice versa.

5.3 Discipline-specific strengths and weaknesses (H3-H5)

DIF patterns. The DIF parameters b_{SO}, b_{PO}, and b_{ED} that resulted from the 27 logistic regressions reported in Table 5 showed the same direction for 20 of 27 items. The correlations were high and significant ($r_{SO,ED}$ = .89, $r_{SO,PO}$ = .88, $r_{ED,PO}$ = .83, *ps* < .001). This supports the hypothesis that disciplines with similar OTL-patterns showed similar DIF-patterns (H3).

Analysis of strengths and weaknesses based on DIF-parameters. For *sociology* students, the DIF-parameters of qualitative items were significantly positive, indicating that these students were more likely to solve qualitative items than expected given their overall test scores. In contrast, no systematic DIF was revealed on quantitative items (see Table 8). Differentiation between the research traditions addressed in the items explained around a third of the variance in DIF-coefficients. Similar results were found for *educational studies*: The DIF-parameters of qualitative items were significantly positive, again indicating a systematic strength in qualitative methods, whereas quantitative items did not show systematic DIF – with research traditions explaining more than a third of the variance in DIF-coefficients (see Table 8). Students of *political science* showed specific strengths or weaknesses neither on qualitative nor on quantitative items. In political science, the differentiation in research traditions explained a much smaller amount of variance (see Table 8).

Table 8: Regression of item-wise DIF-parameters on item characteristics (research traditions addressed in the items) by disciplines (n = 27 per discipline)

	sociology				political science				educational studies			
	b	SE_b	β	$t(24)$	b	SE_b	β	$t(24)$	b	SE_b	β	$t(24)$
quantitative items	-0.31	0.27	-.21	-1.15	-0.32	0.28	-.27	-1.14	-0.39	0.23	-.31	-1.74
qualitative items	0.66	0.27	.45	2.41 *	0.27	0.28	.23	0.99	0.53	0.23	.42	2.36*
R^2_{adj}			0.31				0.13				0.38	
$F(2,24)$			6.42**				2.22				8.26**	

Note. * $p < .05$, ** $p < .01$

Analysis of strengths and weaknesses based on the response behavior on groups of items. Students of sociology, political science and educational studies performed significantly worse on quantitative items compared to psychology students, when controlling for the overall DIF-free RC score (see Table 9). On qualitative items, students of sociology and educational studies significantly outperform psychology students (see Table 9).

Table 9: Performance on groups of items by disciplines

	mean performance on quantitative items (n = 670)				mean performance on qualitative items (n = 671)			
	b	SE_b	β	$t(665)$	b	SE_b	β	$t(666)$
overall DIF-free RC scores	0.09	0.004	.64	22.99***	0.06	0.004	.44	13.67***
sociology	-0.06	0.018	-.11	-3.45***	0.09	0.030	.18	2.99**
political science	-0.06	0.017	-.14	-3.80***	0.03	0.028	.08	1.25
educational studies	-0.08	0.010	-.15	-8.27***	0.08	0.028	.16	2.81**
R^2_{adj}	.46				.21			
F	$F(4, 665) = 155.98$***				$F(4, 666) = 61.00$***			

Note. ** $p<.01$, *** $p<.001$

6. Discussion

6.1 Summary

The objective of this paper was to assess discipline-specific strengths and weaknesses on a newly developed domain-general measure of social-scientific RC. The overall expectation was that students of a discipline systematically outperform other students in groups of items that address content domains highlighted in the respective discipline. The measure has been applied in four social-scientific disciplines alongside a questionnaire assessing OTL.

Moderate DIF was evident in five of 27 items and strong DIF in one item. As expected, given the distinctive methodical and methodological emphases evident in the disciplines' guidelines for curricula, sociology, educational studies and political science showed similar patterns in DIF-parameters. For most of the items, DIF-parameters had the same direction and correlated closely between these three disciplines. This concurred with the disciplines' OTL-patterns: While students of sociology, educational studies and political science reported similar shares of quantitative and qualitative OTL, psychology students differed strongly by reporting a comparatively low share of OTL in qualitative research methods.

The curricular foci evident in OTL-differences coincided with the students' performance. Two approaches have been applied to analyze discipline-specific strengths and weaknesses in qualitative items on the one hand and quantitative items on the other. In the first approach, the DIF-parameters were analyzed for each discipline according to item characteristics – positive DIF-parameters indicated a relative strength of a discipline. In the second approach, the students mean performance on qualitative items and quantitative items were analyzed by discipline while controlling for the students' overall RC scores. Both approaches showed similar results. Students of sociology and educational studies showed significantly positive DIF-parameters for qualitative items, i.e., a relative strength in qualitative methods and significantly outperformed psychology students on these items. Regarding quantitative items, students of sociology, political science and educational studies were significantly outperformed by psychology students. Their DIF-parameters for quantitative items were, however, only insignificantly negative, i.e. no systematic weakness was found. Overall, discipline-specific strengths and weaknesses were found in items addressing research methods highlighted in the respective curricula, as demonstrated by differences in shares of OTL. Hence, the measure adequately maps item-wise ability differences based on OTL reported by the students. Both the ability and OTL-differences are also consistent with the guidelines issued by the professional associations of the disciplines under consideration.

Bundling the items according to the research traditions targeted by the items explained a substantial amount of variance in the DIF-parameters in all disciplines. It explained around one third of the variance in sociology and educational studies and one eighth in political science – students of psychology were used as a reference group. While it has been demonstrated that a multidimensional model differentiating between research traditions did not outperform the unidimensional measurement model of RC (see Section 4.2.1), the DIF-analyses provided evidence of the relevance of these secondary dimensions. If the underlying secondary dimensions that cause DIF were intended to be measured, these are "auxiliary dimensions" (Roussos & Stout, 1996, p. 356) as opposed to "nuisance dimensions" (ibid.) that were not intended to be measured. Regarding the measure at stake, the secondary dimensions (i.e. research traditions) can be interpreted as auxiliary dimensions because the assessment of RC should mirror differences in OTL. If the measure had not mirrored OTL differences, we would have to assume that we had measured a general disposition not addressed in social-scientific research training.

Hence, the interpretation of test scores as a measure of a competency would be questionable as competencies should be sensitive to OTL (Koeppen et al., 2008).

Regarding the *applicability of the proposed RC-measure across disciplines*, there is evidence in support and in refusal. On the one hand, the existence of DIF can be interpreted as a bias of the measure (as it did in the so called first generation of DIF-analyses, see Zumbo, 2007). On the other hand, DIF can be interpreted as benign and even as validation evidence if it is caused by an auxiliary dimension and thus conforms with the test framework (AERA, APA, & NCME, 2014; Roussos & Stout, 1996). Around one third of the variance in DIF parameters could be explained by auxiliary dimensions and indirectly attributed to differences in OTL. This is evidence of the items' validity: They are sensitive to differences in OTL. However, a substantial amount of DIF remained unexplained. It deems worthwhile to further investigate possible causes of the remaining DIF. In further analyses, additional item characteristics should be examined to explain DIF. Promising characteristics are the methodological paradigm addressed (e.g., differentiating items on ethnographic, narrative and phenomenological research or on experimental and survey studies), the types of research projects described in vignettes (e.g., student project, thesis, professional research) or technical criteria (e.g., item difficulty and format). For detailed analyses like these, a larger item pool is needed.

Regarding the construct *student social-scientific research competency* some tentative conclusions can be inferred. Firstly, although the construct is very broad and incorporates different research traditions, it can still be viewed as a unidimensional construct. Secondly, while the findings for scientific discovery indicated domain-generality of the entire construct (Hartmann et al., 2015), the present paper suggested that domain-generality of social-scientific RC is limited to its auxiliary dimensions: Qualitative items seemed to measure the same across disciplines as did quantitative items and items contingent to both research traditions. Thirdly, based on the first two conclusions, social-scientific RC seems to follow a bifactor structure (Reise, 2012): A strong general factor (the unidimensional interpretation) and rather weak group factors representing the research traditions (the auxiliary dimensions) seemed to be present. For the OTL on the other hand, a three-dimensional structure was found that differentiated between the research traditions. Given these findings, a bifactor structure of RC should be further explored. It seems beneficial to address this question using an additional, separate set of items in order to replicate our findings.

6.2 Limitations

The domain-generality of the measure was analyzed comparing the response behavior of students from four disciplines. However, the measure is intended to be applicable across *all* social-scientific disciplines. The findings therefore are limited to these disciplines. For a generalization of findings, replication studies in other disciplines (e.g. ethnology, social work) are needed.

Some conclusions have been drawn on the nature of social-scientific RC. These conclusions were only tentative as the measure at stake was newly developed and the analysis of DIF was based on only 27 units of analysis. Although robust regressions were performed that weighted for outliers, the findings on domain-generality were limited to the item set present. Replication studies on a different set of items are encouraged.

The relation between DIF and OTL was assessed indirectly. Although the disciplines' strength and weaknesses conformed with their emphases in OTL, a causal relation behind this concurrence – despite being probable – cannot be inferred from the analyses presented in this paper. Furthermore, it needs to be pointed out that a convenience sample was used. Thus, the findings on OTL should not be mistaken for conclusions on OTL in these disciplines in general.

7. Conclusion

Overall, the analysis of DIF and the concomitant analysis of differences in OTL proved insightful. The analysis of DIF allowed a comparative perspective, focusing on differences between disciplines instead of commonalities and thus proved useful as a diagnostic tool. The analysis of OTL on the other hand proved useful as an explanatory tool for discipline-specific item-response behavior. By linking curricular foci, differences in OTL and DIF, it was possible to test and prove a chain of coherent hypotheses.

The work presented here adds to the debate on the content of research courses within and between social-scientific disciplines. The analysis of OTL provided evidence on the course contents of different study programs. The newly developed test instrument at hand may additionally provide a first tool to measure RC and initiate a discussion on the objective assessment of RC as a whole. While cross-disciplinary comparisons of ability scores seem premature due to the unexplained variance in DIF, the use of the test for evaluative purposes in the four social-scientific disciplines analyzed is encouraged.

Acknowledgements

This research was supported by a grant from the German Federal Ministry of Education and Research (BMBF), FKZ 01PB14004/B. We thank the graduates and undergraduates for participating in the survey, the lecturers for allotting precious teaching time to our project, and two anonymous reviewers for their helpful comments.

References

Adams, R. J., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: OECD.

AERA – American Educational Research Association, APA – American Psychological Association, & NCME – National Council on Measurement in Education. (2014). *Standards for educational & psychological testing*. Washington, DC: American Educational Research Association.

APSA – American Political Science Association. (2004). *APSA task force on graduate education*. Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED495969.pdf

APA – American Psychological Association. (2011). *APA principles for quality undergraduate education in psychology* (Vol. 54). Washington, DC: APA.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Blooms taxonomy of educational objectives*. New York, NY: Longman.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*(1), 3–13. http://doi.org/10.1027/2151-2604/a000194

Blömeke, S., Suhl, U., & Döhrmann, M. (2013). Assessing strengths and weaknesses of teacher knowledge in Asia, Eastern Europe, and Western Countries: Differential item functioning in Teds-M. *International Journal of Science and Mathematics Education*, *11*(4), 795–817. http://doi.org/10.1007/s10763-013-9413-0

Borg, S. (2010). Language teacher research engagement. *Language Teaching*, *43*(4), 391–429. http://dx.doi.org/10.1017/S0261444810000170

British Academy. (2012). *Society counts: Quantitative skills in the social sciences*. London, England. Retrieved from http://www.britac.ac.uk/sites/default/files/BA Position Statement – Society Counts.pdf

Butler, D. (2006). Frames of inquiry in educational psychology: Beyond the quantitative-qualitative divide. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (pp. 903–927). Mahwah, NJ: Lawrence Erlbaum.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

DGFE – Deutsche Gesellschaft für Erziehungswissenschaft. (2004*). Kerncurriculum für das Hauptfachstudium Erziehungswissenschaft*. Retrieved from http://www.dgfe.de/fileadmin/OrdnerRedakteure/Stellungnahmen/2004_01_KC_HF_EW.pdf

Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teaching in Higher Education*, *19*(3), 242–253. http://doi.org/10.1080/13562517.2013.860105

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *2*(3), 28–45.

Gess, C., Geiger, C., & Ziegler, M. (in press). Social-scientific research competency: Validation of test score interpretations for evaluative purposes in higher education. *European Journal of Psychological Assessment*.

Groß Ophoff, J., Schladitz, S., Leuders, J., Leuders, T., & Wirtz, M. A. (2015). Assessing the development of educational research literacy: The effect of courses on research methods in studies of educational science. *Peabody Journal of Education*, *90*(4), 560–573. http://doi.org/10.1080/0161956X.2015.1068085

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314–329. http://doi.org/10.1111/j.1745-3984.2009.00083.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.

Harrell, F. E. J. (2015). *rms: Regression modeling strategies*. Retrieved from: https://CRAN.R-project.org/package=rms

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2-3), 57–63. http://doi.org/10.1016/j.stueduc.2009.10.002

Hartmann, S., Upmeier zu Belzen, A., Krüger, D., & Pant, H. A. (2015). Scientific reasoning in higher education. *Zeitschrift für Psychologie*, *223*(1), 47–53. http://doi.org/10.1027/2151-2604/a000199

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hussy, W., Schreier, M., & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften – für Bachelor*. Berlin, Germany: Springer Medizin.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. http://doi.org/10.1207/S15324818AME1404_2

Katz, I. R. (2007). Testing information literacy in digital environments: ETS's iSkills assessment. *Information Technology and Libraries*, *26*(3), 3–12.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education*, *16*(3), 385–402. http://doi.org/10.1007/BF03173189

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie*, *216*(2), 61–73. http://doi.org/10.1027/0044-3409.216.2.61

Kopf, J., Zeileis, A., & Strobl, C. (2015a). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103. http://doi.org/10.1177/0146621614544195

Kopf, J., Zeileis, A., & Strobl, C. (2015b). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. http://doi.org/10.1177/0013164414529792

Lettau, A., & Breuer, F. (2007). Forscher/innen-Reflexivität und qualitative sozialwissenschaftliche Methodik in der Psychologie. *Journal für Psychologie*, *15*(2).

Lutz, W., & Knox, S. (2014). *Quantitative and qualitative methods in psychotherapy research*. New York, NY: Routledge.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research & Perspective, 11*(3), 71–101. http://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2(n) contingency tables: A unified framework. *Journal of the American Statistical Association*, *100(471)*, 1009–1020. http://doi.org/10.1198/016214504000002069

McKinney, K., Howery, C. B., Strand, K. J., Kain, E. L., & Berheide, C. W. (2004). *Liberal learning and the sociology major updated: Meeting the challenge of teaching sociology in the twenty-first century*. Washington, DC: American Sociological Association.

Mruck, K., & Mey, G. (2000). Qualitative Forschung. In F. Jacobi & A. Poldrack (Eds.), *Klinisch-psychologische Forschung: ein Praxishandbuch* (pp. 191–208). Göttingen, Germany: Hogrefe.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Peden, B. F., & Carroll, D. W. (2009). Historical trends in teaching research methods by psychologists in the United States. In M. Garner, C. Wagner, & B. Kawulich (Eds.), *Teaching Research Methods in the Social Sciences* (pp. 23–34). Farnham, United Kingdom: Ashgate.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. http://doi.org/10.1080/00273171.2012.715555

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Eds.), *Handbook of methods for detecting test bias* (pp. 9–30). Baltimore, MD: Johns Hopkins University Press.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *The Journal of Applied Psychology*, *91*(6), 1292–1306. http://doi.org/10.1037/0021-9010.91.6.1292

Stone, A. (2006). *A psychometric analysis of the statistics concept inventory* (Doctoral dissertation). Retrieved from http://www.shareok.org/bitstream/handle/11244/1013/3208004.PDF?sequence=1

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York, NY: Springer.

Wagner, C., Garner, M., & Kawulich, B. (2011). The state of the art of teaching research methods in the social sciences: Towards a pedagogical culture. *Studies in Higher Education*, *36*(1), 75–88. http://doi.org/10.1080/03075070903452594

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wissenschaftsrat. (2006). *Empfehlungen zur künftigen Rolle der Universitäten im Wissenschaftssystem*. Berlin, Germany. Retrieved from https://www.wissenschaftsrat.de/download/archiv/7067-06.pdf

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., & Kopf, J. (2016). *Psychotools: Infrastructure for psychometric modeling*. *R package version 0.4-2*. Retrieved from https://CRAN.R-project.org/package=psychotools

Ziegler, M., Kemper, C. J., & Lenzner, T. (2015). The issue of fuzzy concepts in test construction and possible remedies. *European Journal of Psychological Assessment*, *31*(1), 1–4. http://doi.org/10.1027/1015-5759/a000255

Zumbo, B. 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, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. http://doi.org/10.1080/15434300701375832

Annex A.

Item A1: QN_ML_3

Research tradition	*Level*	*Correct answer*
quantitative methodology	remember knowledge	A

What does it mean, if a significant effect is found on a significance level of 5 percent?

(Please select one option only)

❍ A	The probability of obtaining this or an even stronger effect — although actually no effect exists — is at most 5 percent.
❍ B	The probability of obtaining no effect — although actually an effect exists — is at most 5 percent.
❍ C	The probability that an existing effect can actually be found is at least 95 percent.
❍ D	An effect can only be found, if at least 5 percent of the respondents deviate from the confidence interval.

Item A2: QL_MD_4

Research tradition	*Level*	*Correct answer*
qualitative methods	apply knowledge	B2

Research project on behalf of the German association for inland water transportation (duration: 2 years)

Research topics: values and norms of inland water navigators

State of research: So far, only few studies on this topic exist; theories are rather vague, i.e. theories do not explain relationships and circumstances

Research design:
- methodology: Grounded Theory (qualitative methodology)
- realization of 10 narrative interviews
- development of a theory on values and norms of inland water navigators based on the interview transcripts

It has to be decided, when and according to which criteria the selection of potential interview partners should be performed. What decision should the researcher make in this research project?
(Please select one option each, i.e. select two options in total)

Criteria of selection:

If possible, the researcher should select the interview partners ...	
❍ A	based on a random selection.
❍ B	based on theoretical considerations.

Moment of selection:

If possible, the researcher should select the interview partners ...	
❍ 1	in advance, i.e. before the first interview.
❍ 2	iteratively, i.e. during data collection and data analysis

Item A3: Item BM_RP_9

Research tradition	*Level*	*Correct answer*
both research traditions	make judgments	D

Bachelor thesis

Research question: „Which procedures of social control do squatters make use of?"

Research design:
- survey study with 100 squatters (survey was validated in several studies)
- interviews with 4 squatters
- previous discussion of the questionnaires in a colloquium

Research results:
- The findings from the quantitative survey study and from the qualitative interviews are conflicting: indications for social control could be found in interviews but not in the survey study.

In the bachelor thesis the conflicting findings were discussed equally. Reasons for the inconsistencies were not discussed.

What is the biggest problem of this research project?
(Please select one option only)

❍ A	Only the quantitative findings should have been discussed in the bachelor thesis.
❍ B	Only the qualitative findings should have been discussed in the bachelor thesis.
❍ C	Reasons for inconsistencies should have been analyzed in an additional interview study and an extension of time for submission of the thesis requested.
❍ D	Reasons for inconsistencies should have been discussed with the thesis advisor or with fellow students and described and discussed in the bachelor thesis.

Annex B.

Table B1: Translation of items used in OTL-Scales (n=309)

	Item	*M*	*SD*	Discr.[a]
Quantitative OTL	Fundamental concepts (falsification, hypothesis, type I-error, type II-error, etc.)	.94	.23	.52
	Quality criteria in research (reliability, validity, etc.)	.93	.25	.43
	Quantitative research designs (experimental, quasi-experimental, laboratory study, field study, etc.)	.85	.36	.52
	Measurement (scales, measurement levels, etc.)	.94	.25	.49
	Development of surveys (formulating questions and answers, ordering of questions, sources of bias, etc.)	.78	.42	.38
	Drawing a sample (representativity, random sample, cluster sample, stratified sample, etc.)	.82	.38	.44
	Descriptive statistics (distribution, mean/median, variance, histogram, boxplot, scatter plot, etc.)	.91	.29	.56
	Bivariate statistics (crosstab, chi-square-test, correlation, t-test, bivariate regression, etc.)	.85	.36	.52
	Multivariate statistics (linear regression, logistic regression, analysis of variance, etc.)	.72	.45	.51
	Advanced statistics (longitudinal, multilevel, factor, cluster analysis, structural equation modeling, etc.)	.27	.45	.32
	Test theory (classical test theory, item response theory, latent trait theory, scaling, etc.)	.47	.50	.47
Qualitative OTL	Qualitative research designs (case study, analysis of documents, classification, comparison study, etc.)	.51	.50	.59
	Sampling (theoretical sampling, snowball-sampling, etc.)	.42	.49	.44
	Interview (types of interviews, types of questions, interview guidelines)	.84	.37	.49
	Group discussion (group interview, -discussion, grouping, etc.)	.41	.49	.56
	Participatory observation (observation protocol, videography, phonogram, etc.)	.62	.49	.51
	Archiving data (protocol, transcription, transcription abbreviations, etc.)	.45	.50	.57
	Qualitative content analysis (Mayring, Kuckartz, etc.)	.38	.49	.52
	Grounded-Theory (Glaser, Strauss, Charmaz, etc.)	.31	.46	.61
	Narration analysis (Schütze, Matthes, etc.)	.21	.41	.56
	Objective hermeneutics (Oevermann, Schneider, etc.)	.26	.44	.52
	Documentary method (Bohnsack, Przyborski, etc.)	.15	.36	.47
Research process OTL	How to find a good research question (finding a topic, formulating a question, specifying, etc.)	.73	.45	.51
	How to plan a study (assessing the state of research, defining the key concepts, etc.)	.71	.45	.54
	How to assess the state of research (literature research, excerpt, summarizing, etc.)	.74	.44	.54
	How to consider practical obstacles (entering a field, time planning, privacy concerns etc.)	.57	.50	.51
	How to write scientifically (planning, structuring, citing, publishing, etc.)	.91	.29	.29

Note. [a] Discr. is the item-total-correlation using part-whole correction (correlation between item and total score of all items except the respective item analyzed).