Groß Ophoff, Jana; Wolf, Raffaela; Schladitz, Sandra; Wirtz, Markus

Assessment of educational research literacy in higher education: Construct validation of the factorial structure of an assessment instrument comparing different treatments of omitted responses

Journal for educational research online 9 (2017) 2, S. 37-68

Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 9 (2017), No. 2, 37–68

Jana Groß Ophoff, Raffaela Wolf, Sandra Schladitz & Markus Wirtz

Assessment of Educational Research Literacy in Higher Education: Construct validation of the factorial structure of an assessment instrument comparing different treatments of omitted responses

Abstract

The ability to purposefully access, reflect, and use evidence from educational research (Educational Research Literacy) are key competencies of future professionals in educational practice. A test instrument was developed to assess Educational Research Literacy with the competence facets Information Literacy, Statistical Literacy, and Evidence-based Reasoning. Even though there are certain overlaps with generic concepts like critical thinking or problem solving, Educational Research Literacy is acquired within its reference disciplines. This contribution aimed to delve deeper into the question which factorial model is most appropriate. Four competing models were compared: unidimensional, three-dimensional, and two bifactor models. The comparison was based on a study of 1360 students at six German universities and was validated by another study of 753 students at three universities. The results also were examined relative to the scoring of omitted responses and the booklet design used in the first study. The results indicate that the four-dimensional bifactor model was the most appropriate: Educational Research Literacy seems to consist of one dominant factor and three secondary factors. The results also support handling both omitted and not-reached responses as missing information. Subsequently, the results are critically discussed rela-

Dr. Jana Groß Ophoff (corresponding author), Institute for Educational Studies, University for Education Freiburg, Kunzenweg 21, 79117 Freiburg, Germany
e-mail: jana.grossophoff@ph-freiburg.de

Dr. Raffaela Wolf, Lexia Learning, a Rosetta Stone Company, 300 Baker Avenue, Suite 320, Concord Massachusetts 01742, United States of America
e-mail: rwolf@lexialearning.com

Sandra Schladitz, Institute for Psychology, Department of Psychology, Philipps-University Marburg, Gutenbergstraße 18, 35032 Marburg, Germany
e-mail: sandra.schladitz@staff.uni-marburg.de

Markus Wirtz, Institute for Psychology, University of Education Freiburg, Kunzenweg 21, 79117 Freiburg, Germany
e-mail: markus.wirtz@ph-freiburg.de

tive to the requirements for assessing and for imparting competencies in higher education. Recommendations for future research are stated.

Keywords
Educational Research Literacy, Higher education, Competency tests, Dimensional analysis, Missing data

Erfassung bildungswissenschaftlicher Forschungskompetenz in der Hochschulbildung: Konstruktvalidierung der Faktorstruktur eines Testverfahrens unter Berücksichtigung des unterschiedlichen Umgangs mit ausgelassenen Antworten

Zusammenfassung
Evidenz aus bildungswissenschaftlicher Forschung zielgerichtet erschließen, reflektieren und anwenden zu können (sog. Bildungswissenschaftliche Forschungskompetenz, BFK) ist zentral für Fachpersonal im Bildungswesen. Zur Erfassung dieser Kompetenz (mit den Facetten Informationskompetenz, Statistische Kompetenz, Evidenzbasiertes Schlussfolgern) wurde ein Testinstrument entwickelt. Trotz Gemeinsamkeiten mit generischen Konzepten wie kritisches Denken oder Problemlösen wird BFK innerhalb der Bezugsdisziplinen erworben und entwickelt. Dieser Beitrag widmet sich der Frage nach dem am besten passenden Strukturmodell. Hierzu wurden ein eindimensionales Modell, ein dreidimensionales Modell und zwei bifaktorielle Modelle verglichen. Der Modellvergleich basierte auf Daten einer Studie an sechs deutschen Hochschulen (1360 Studierende) und wurde anhand einer Folgestudie an drei Hochschulen validiert (753 Studierende). Untersucht wurden auch Unterschiede bezüglich der Kodierung ausgelassener Antworten oder dem Testheftdesign der ersten Studie. Die Ergebnisse sprechen für das vierdimensionale bifaktorielle Modell, wonach BFK aus einem dominanten Faktor und drei Sekundärfaktoren besteht. Die Ergebnisse unterstützen die Empfehlung, Auslassungen als fehlende Information in den Analysen zu belassen. Die Ergebnisse werden abschließend hinsichtlich der Anforderungen an die Erfassung und Vermittlung von Kompetenzen im Hochschulsektor diskutiert und Desiderata für künftige Forschung benannt.

Schlüsselwörter
Bildungswissenschaftliche Forschungskompetenz; Hochschulbildung; Kompetenztests; Dimensionale Analyse; Bifaktorielle Modelle; Fehlende Daten

1. Relevance of Educational Research Literacy

Educational Research Literacy (ERL) is the ability to purposefully access, comprehend, and reflect scientific information as well as apply the resulting conclusions to problems with respect to educational decisions (Groß Ophoff, Schladitz, Lohrmann, & Wirtz, 2014; McMillan & Schumacher, 2010; Shank & Brown, 2007). ERL can be described as part of Assessment Literacy (Brookhart, 2011; DeLuca, LaPointe-McEwan, & Luhanga, 2016), comprised of different competence facets like Information Literacy (e.g., Catts & Lau, 2008), Statistical Literacy (e.g., Ben-Zvi & Garfield, 2004), and Critical Thinking (e.g., Meltzoff, 2010). These facets can be allocated to the research cycle, which was used in the current study as conceptual framework for the development and construct validation of an assessment of ERL in Higher Education in the present study (see section 2).

Due to continued scientific progress, advanced ERL is important not only for social participation (cf. Grundmann & Stehr, 2012), but is a fundamental requirement for Continuing Professional Development (e.g., Jindal-Snape, Hannah, Smith, Barrow, & Kerr, 2009; Rankin & Becker, 2006). However, Borg (2010) emphasized, that although current and future practitioners in education need to engage themselves *with research*, they do not necessarily have to engage themselves *in* research. Nonetheless, engagement with research in educational contexts is not without difficulties. While scientific evidence is formulated falsifiable and generalizable, educational practice aims at solving problems instantly and efficiently. It is this gap between theory and practice that frequently leads both students and practitioners to view research information as abstract, irrelevant factual knowledge, which cannot be applied to practical problems (Benson & Blackman, 2003; G. T. L. Brown, 2004; Hammersley, 2004; Harper, Gannon, & Robinson, 2012; Zeuch, Förster, & Souvignier, 2017). Furthermore, the ability to reflect and use evidence is neither necessarily developed nor retrieved optimally in adulthood (Barchfeld & Sodian, 2009). As students, graduates and professionals will be responsible for imparting relevant competencies to future generations, education plays a central role. Hence, future educators must be trained to use research knowledge in practice (Shank & Brown, 2007). Higher Education institutions particularly are suitable for this as they provide research-based education.

Research literacy currently is included in the general definitions of standards and objectives for German Higher Education degrees (Standing Conference of the Ministers of Education and Cultural Affairs, 2005; German Science Council, 2000), and can also be found in degree programs in Educational Science, e.g., in Teacher Education curricula (Ministry of Cultural Affairs of Baden-Württemberg, 2011; Standing Conference of the Ministers of Education and Cultural Affairs, 2004). In German Higher Education, Educational Science is an umbrella term for different study programs that address the theory and practice of education and training, both from a more general view (e.g., Teacher Training, Educational Studies[1])

1 German: Erziehungswissenschaft

and with focus on certain age groups (e.g., Early Education) or specialized subjects (e.g., Health Education).

Traditionally, German Higher Education institutions offered one-tier study programs that led to Diplom- or Magister Artium degrees or were completed, for example in the case of teacher training, by the so-called State Examination. Following the Bologna Reform agreement in 1999, however, Germany has committed to switch over to the Bachelor and Master degree system by 2020, which has mostly been completed as of 2011 (Federal Ministry of Education and Research, 2015). But in Teacher Education, only 11 of the 16 German federal states have implemented the two-tier degree system as of 2015 (Standing Conference of the Ministers of Educations and Cultural Affairs, 2015). Blömeke and Zlatkin-Troitschanskaia (2013) emphasized that the ongoing reorganization and change processes in the German heterogeneous tertiary sector require a theoretical and empirical foundation for developing and implementing sustainable measures for quality assurance and development. The investigation presented in this paper draws on this point by developing and validating a test instrument for the assessment of ERL (Groß Ophoff et al., 2014), which is intended to be used for measurement and evaluation on the student, course, or institutional level. However, modeling and assessing the development and effects of academic competencies and their influencing factors with validity and reliability relies heavily on research methodology (Blömeke, Gustafsson, & Shavelson, 2015). For example, the test performance in studies, such as the one presented here, has no consequences for participating students (Cole, Bergin, & Whittaker, 2008). These so-called low-stakes tests (unlike university exams), therefore, entail both a low willingness to participate and low test-taking efforts, with the latter typically reflected in the proportion of omitted responses (Köhler, Pohl, & Carstensen, 2015; Wise & DeMars, 2005). Even though omissions are quite common in psychological and educational research (Lüdtke, Robitzsch, Trautwein, & Köller, 2007), missing data can lead to biased parameter estimates, and ultimately to inaccurate conclusions (Durrant, 2005; Peugh & Enders, 2004; Schafer & Graham, 2002; Wirtz, 2004).

2. Conceptual framework

According to Davies (1999), educational professionals at all levels should be able (a) to pose answerable questions; (b) search for relevant information; (c) read and critically appraise evidence; and (d) evaluate and (e) apply the resulting conclusions to their educational needs and environments. These requirements correspond to the steps of the abovementioned research cycle. Comparable process descriptions can be found in theoretical models, too, in which learning is described as an evidence-based process to construct new knowledge (e.g., Davidson, 2013; Pedaste et al., 2015). Some curricular models (e.g., Calzada Prado & Marzal, 2013; Mandinach & Gummer, 2016) use the research cycle to structure learning objec-

tives and to differentiate performance levels of ERL. For example, Willison and O'Regan (2007) described the development of ERL throughout the course of study as the progression from mere conception to application of research information (in the sense of research literacy), and, eventually, unsupported implementation of research (in the sense of research competency, cf. Gess, Wessels, & Blömeke, in this issue).

Overall, ERL typically is assessed based on self-reports (Adedokun, Bessenbacher, Parker, Kirkham, & Burgess, 2013; Borg & Alshumaimeri, 2012; Braun, Gusy, Leidner, & Hannover, 2008; Ntuli & Kyei-Blankson, 2016), but correlations between subjective and objective competency measures are usually low (Lowman & Williams, 1987; Norris, Phillips, & Korpan, 2003; Schladitz, Groß Ophoff, & Wirtz, 2015). Empirical approaches via assignment of test instruments in the education sector can be found, but still are scarce and psychometrically weak (e.g., Reeves & Honig, 2015; cf. Gotch & French, 2014). This is not the case in the field of evidence-based medicine. For example, Shaneyfelt et al. (2006) organized their review of Evidence-Based Practice teaching evaluation instruments in accordance to the abovementioned research steps.

Depending on the objectives linked to a specific problem, the research steps likely are realized in different ways: If there is a need to gain a better understanding of a problem (in terms of research methodology: theory building, e.g., Colquitt & Zapata-Phelan, 2007; Wirtz & Strohmer, 2016), it is to be expected that the research steps will be broader in scope and rather inductive. For example, a teacher may perceive the constant disruptive behavior of a particular student as problematic. To identify the causes, he or she may utilize an inductive empirical approach by seeking dialogue with the parents. If, however, the available information about determining factors can be considered as sufficient from the educator's perspective, hypothetical-deductive methods are more appropriate. Hence, the focus of the approach probably will be more focused to identify, apply, and evaluate appropriate interventions (e.g., inclusion of a school social worker).

Evidence on certain facets of ERL can be found in educational research and related fields, but different aspects of the research cycle are emphasized due to discipline-specific focuses. For example, the ability to formulate appropriate (research) questions and to search and evaluate necessary information – which corresponds to the first (a) and second (b) research step – usually is investigated under the term *Information Literacy* (IL) in information science (e.g., Blixrud, 2003). Moving from information search to reflection as the subsequent third step (c), it is necessary to be able to read and organize data, and interact with different representations. This ability to search and evaluate especially numerical information is investigated as *Statistical Literacy* (SL) in the field of mathematics education (e.g., Groth, 2007; Rott, Leuders, & Stahl, 2015; Watson & Callingham, 2003) or – with a more prominent research-methodological focus – psychology education (Schweizer, Steinwascher, Moosbrugger, & Reiss, 2011). The fourth step (d) requires the ability to substantiate reasoning or critically evaluate given conclusions with respect to scientific quality criteria, which is referred as *Evidence-Based*

Reasoning (ER) hereafter. Corresponding research approaches can be found in research on Science Literacy (STEM education, e.g., N. J. S. Brown, Nagashima, Fu, Timms, & Wilson, 2010; D. Kuhn, Iordanou, Pease, & Wirkala, 2008) or on Critical Thinking (psychology education, e.g., Dunn, Halonen, & Smith, 2008; Lawson, 1999). The fifth (e) and final step of integrating multiple sources of evidence to make logical decisions and identifying unresolved and future research questions is, among others, addressed by research on *Problem Solving* (e.g., Novick & Bassok, 2005; Phye, 2001).

In the field of competency assessment, psychometrically sound test instruments provide an opportunity for criterion-referenced interpretation of underlying models, which can both stimulate curriculum development and facilitate feedback about learning goals and gains (Hartig, 2008; Wilson & Scalise, 2006). According to Prenzel, Walter, and Frey (2007), probabilistic test theory, the basis for the reported analyses in this paper, permits to validate theoretically plausible assumptions about the the dimensional structure of a construct (e.g., by comparing competing models, cf. Adams, Wilson, & Wang, 1997). Thus, a construct valid measurement can be assumed when there is empirical evidence that supports (a) a logistic association of item responses and the according underlying latent trait and (b) the hypothesized correlational structure within and between constructs (cf. Cronbach & Meehl, 1955; Newton & Shaw, 2014). For example, Kretzschmar, Neubert, Wüstenberg, and Greiff (2016) showed that complex problem solving, a concept adjacent to ERL, represents unique variance that is not accounted for by intelligence. With respect to ERL, Schladitz et al. (2015) reported that this competence is related to, but distinguishable from fluid intelligence, too (i.e., convergent validity). The aforementioned need for research applies especially to evidence about the structure of ERL, but a few examples based on objective tests can be found. Gotch and French (2013) described measurement knowledge (as indicator of Assessment Literacy) as a one-dimensional model, but without comparison to competing multidimensional models. Similarly, assessment of SL (Watson & Callingham, 2003) and IL (O'Connor, Radcliff, & Gedeon, 2002) were described as unidimensional construct without comparisons to multidimensional models. Based on the self-assessment of doctoral students and candidates for scientific degrees of different study programs, Olehnovica, Bolgzda, and Kravale-Pauliņa (2015) identified three research competency facets: informative, communicative, instrumental. Although the focus is on *engagement in research* (Borg, 2010), the concepts of informative and instrumental ability show some similarities with the conceptual framework of ERL. Informative research competency refers to the competence facet IL, and instrumental competency describes the overall ability to move through the research cycle.

Research, however, has grown in recent years (e.g., Schmid, Richter, Berthold, Bruns, & von der Mühlen, 2013; Trempler, 2013) – not least because of the funding initiative *Modeling and Measuring Competencies in Higher Education* (KoKoHs) by the German Federal Ministry of Education and Research (Blömeke & Zlatkin-Troitschanskaia, 2013). Within this initiative, the joint project *Learning*

the Science of Education (LeScEd) also aims for the theory-based conceptualization and empirical validation of a comprehensive ERL model (Schladitz et al., 2013). Based on preliminary analyses in which both omitted and not-reached responses were treated as missing data, Groß Ophoff et al. (2014) introduced evidence that suggests a one-dimensional model of ERL. But after recoding omitted responses as incorrect, a three-dimensional Rasch model with the subdimensions IL, SL, and ER was identified as the best fitting compared both to the less parsimonous two-parameter logistic (2PL) model and other competing, theoretically plausible models (two-dimensional: research steps; three-dimensional: cognitive requirements). Analysis in a small subsequent study of student development during courses on research methods in educational science was based on this three-dimensional model (Groß Ophoff, Schladitz, Leuders, Leuders, & Wirtz, 2015). The competence facets, however, were highly intercorrelated ($r \geq .68$), which could indicate a general underlying factor (Reise & Revicki, 2014). According to Reise, Moore, and Haviland (2010), it is not uncommon that item response data appears consistent with both unidimensional and multidimensional latent structures.

In summary, the current analyses not only aim to delve deeper into the question of which factorial structure is most appropriate for the given test instrument, but to solve the apparent structural ambiguity – with special attention to the effect of different treatments of omitted responses. For this purpose, another plausible interpretation will be considered in which ERL consists of one dominant factor representing the generic aspect (G) of ERL, and secondary factors of IL, SL and ER representing specific aspects in relation to the requirements of the research cycle (i.e., bifactor model, cf. Holzinger & Swineford, 1937). This is related to the issue of whether ERL can be understood as generic ability. Although there are certain overlaps with concepts like academic skills (Clanchy & Ballard, 1995) or so-called key competencies like critical thinking (D. Kuhn, 1999) or problem solving (Mayer & Wittrock, 2006), ERL is acquired within and influenced by its reference disciplines, and can be seen – at least in part – as a domain-specific ability (Lea & Street, 2006; Wecker, Hetmanek, & Fischer, 2014). The results from the presented study, therefore, may have implications for the curricular alignment and structure of imparting ERL to students of Educational Science, too.

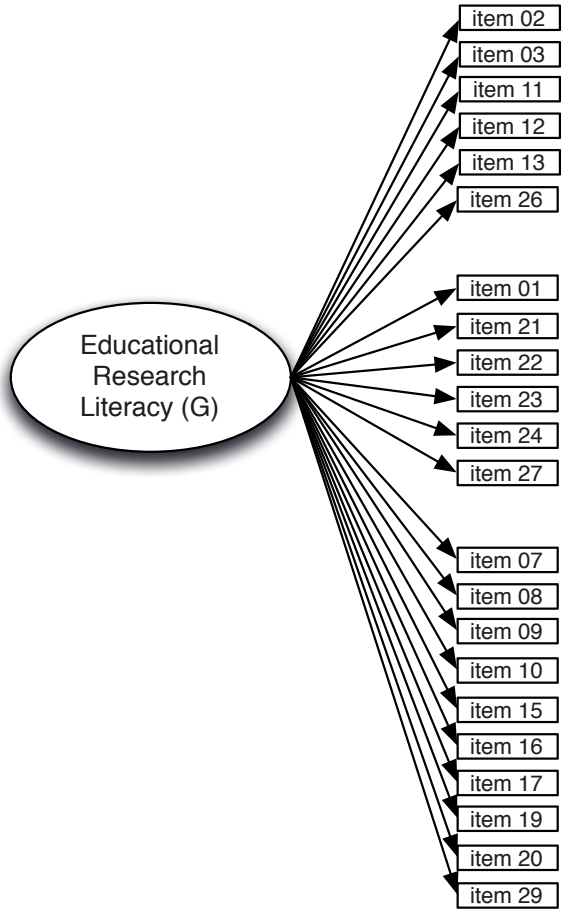
3. Models

The one- and three-factorial models were evaluated and contrasted to two different bifactor models – both for the treatment of omitted responses as ignorable missing data (*condition a*, see section 4.3) and the treatment of omitted responses as incorrect (*condition b*). The reported analyses are based on data from the main study (*study 1*: winter semester 2012/2013/summer semester 2013; cf. Groß Ophoff et al., 2014) and are contrasted to data from the first assessment time point of another subsequent study (*study 2:* summer semester 2014). This comparison aimed to

investigate further whether the assumed factorial structure is a sample-specific result or can be generalized to other independent samples. To determine if the assumed factorial structure actually was an artifact of the treatment of omitted responses or even the assessment procedure itself (e.g., testlet effect, c.f. Wainer & Kiely, 1987), another bifactor model was considered. In summary, the following competing models assuming different structural components of ERL were defined and analyzed:

- Model 1: ERL is assumed to be a one-dimensional ability that covers the requirements of the whole research cycle (model 1, see Figure 1)
- Model 2: ERL is assumed as multidimensional ability, which is composed of three subdimensions (model 2, see Figure 2): the ability to outline and exploit a problem space with appropriate search strategies (IL), the ability to reflect mathematical-statistical representations of evidence (SL), and the ability to critically evaluate evidence-based argumentation and reasoning (ER)

Figure 1: One-dimensional model (model 1)



- Model 3: essentially combines the one- and the three-dimensional model in a bi-factor model with the generic aspect (G) of ERL and secondary factors of IL, SL and ER (see Figure 3)
- Model 4: characterizes another bifactor structure with a general latent factor (G) from 20 secondary factors representing the booklet design from study 1 (for illustration see Figure 6).

Figure 2: Between-item multidimensional model (model 2)

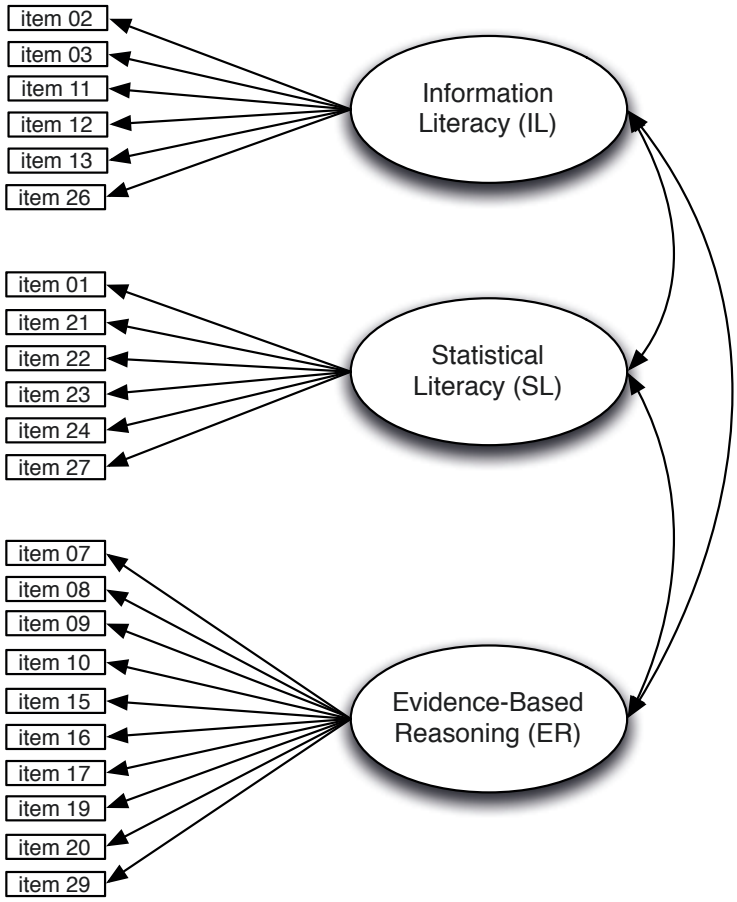
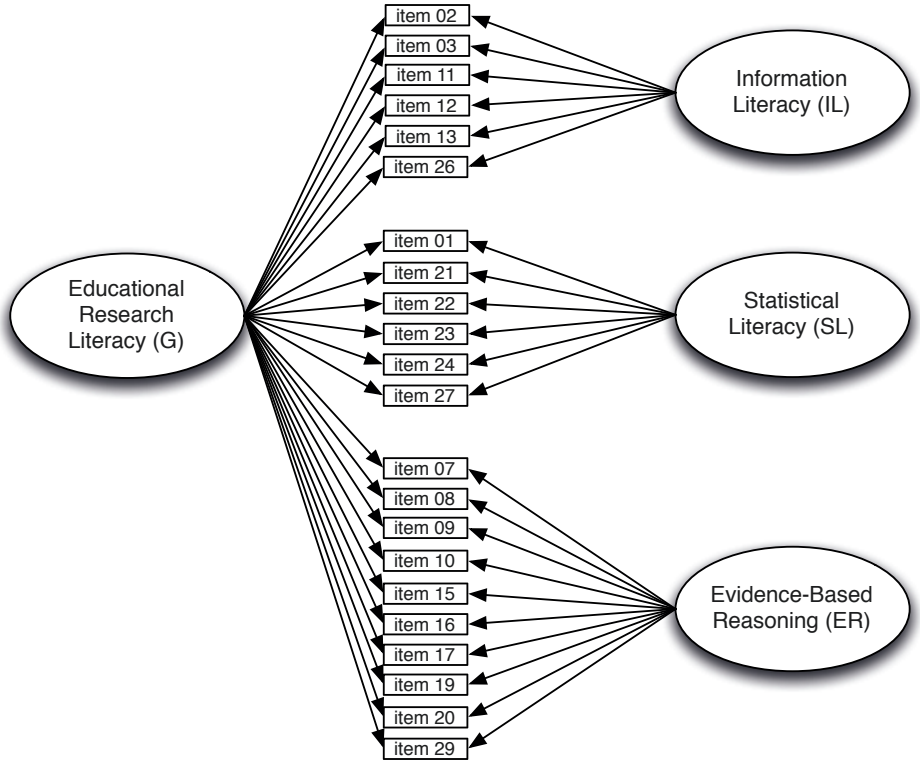


Figure 3: Bifactor Model (model 3)



4. Methods

4.1 Data collection and samples

Analyses were conducted utilizing data sets from two studies: the first from a (large-scale) study aimed at item generation and test standardization (study 1: winter semester 2012/2013 and summer semester 2013; cf. Groß Ophoff et al., 2014). And the second from a subsequent longitudinal study, which was conducted to capture individual learning gains in ERL over the course of one semester (study 2*:* summer semester 2014). For this paper, the sample from the first assessment time point at the start of the semester (study 2) was used to validate the results from study 1. In both studies, participants were recruited upon request in lectures. Participation was voluntary and anonymous. To ensure standardized implementation, test administrators conducted the tests.

In study 1, 1360 students of Educational Science at six German universities[2] were recruited, and 753 students from three universities[3] were recruited in study 2 (see Table 1). The samples were not statistically different in age or the percentage of women. There was a statistically significant difference in the average grade of university entrance qualification (Abitur), but the effect was negligible ($F_{(.05;\ 1)} = 7.447$; $\eta_p^2 = .004$). Teacher Training students represented the largest group, followed by Educational Studies students, and then other study programs (e.g., Early Education, Health Education, Educational Psychology) with the latter accounting for less than 10 % in each sample.

Table 1: Descriptive statistics

	Study 1	Study 2
n	1360	753
Age, *M* (*SD*)	22.9 (3.95)	22.8 (3.97)
Gender (% female)	75.9%	78.2%
Average grade Abitur*, *M (SD)*	2.4 (0.57)	2.3 (0.60)
Study program (first two most frequent)	62%Teacher Training 23% Educational Studies	51% Teacher Training 24% Educational Studies

Note. Study 1: winter semester 2012/2013 and summer semester 2013. Study 2: summer semester 2014.

n = number of study participants; *M (SD)* = mean (standard deviation). *Abitur = German University Entrance Qualification, grades range from 1 to 6 (4 as lowest passing grade) with lower numbers indicating better results.

4.2 Test instrument and booklet design

The conceptual framework described above was used to develop a test instrument for assessing ERL in Higher Education. During the first half of the research program, compiling an extensive item pool was paramount. For this purpose, new test items were generated and already published test items (a.o. Heinze, 2008; McMillan & Schumacher, 2010; Watson & Callingham, 2003) were translated and adapted to educational topics. To optimize the content validity of early drafts, experts on educational research (post-doctoral level or higher) reviewed the material. In addition to concrete suggestions for improvement, it was recommended to focus on forced-choice items that are more easily scored than open-ended tasks. For the same reasons, the development of test items for the competence facet *Problem Solving* (5th research step, see section 2) was postponed, because it usual-

2 Study 1: University of Education Freiburg, Albert-Ludwigs University Freiburg, University Koblenz-Landau, University Göttingen, Free University Berlin, University Duisburg-Essen

3 Study 2: University of Education Freiburg, Albert-Ludwigs University Freiburg, University Koblenz-Landau

ly is assessed via complex, text-intensive performance-tasks (cf. Collegiate Learning Assessment; Klein, Benjamin, Shavelson, & Bolus, 2007).

In another preliminary study, test persons (five undergraduate students, one PhD student) were asked to think aloud while working on selected tasks, and further evidence on understandability and solvability of the test items was gained. After final revision, more than 200 test items, most in forced-choice format, were available for the standardization study (study 1).

Figure 4: Test item for the competence facet Information Literacy. The correct solutions are checked

In order to combine keywords for database search, the logical operators AND, OR, and NOT can be used.
Each of them leads to different search results.

Assign the keyword combination to the referring question.
(Multiple selection is possible)

	Heterogeneity AND Elementary School	Heterogeneity OR Elementary School	Heterogeneity NOT Elementary School
a) Is it possible to compensate for heterogeneity of elementary school children?	☒	☒	☐
b) Is it possible to compensate for migration-related disparities in learning conditions of high school students?	☐	☐	☒
c) Is it possible to compensate for heterogeneity in learning conditions in secondary education?	☐	☐	☒

Each item was assigned to one ERL facet according to the conceptual framework. IL items mainly focused on search strategies for problem-specific research information (translated example see Figure 4), the comprehension of different types of academic documents, or the formulation of adequate research questions. In the given example, semantically correct keyword combinations (front row) needed to be identified in relation to particular research questions (left column). The questions (terminology included) were inspired by an original research article (Kopp & Martschinke, 2011), and addressed a typical research topic in Educational Science. In order to account for possible differences in prior knowledge, a short introductory note about Boolean operators in database search was provided (see Figure 4, grey box on top). Items for the competence facet SL usually require analysis and interpretation of descriptive statistics (e.g., tables, figures, short textual reports),

which are common components of scholarly articles (McMillan & Schumacher, 2010; Shank & Brown, 2007). The item stems were based on published findings from empirical educational research (e.g., PISA 2009; Naumann, Artelt, Schneider, & Stanat, 2010) or fictitious examples relating to educational practice (e.g., school internal teacher survey, class results in mandatory school performance test). Tasks for the facet ER typically consist of two research abstracts, which were based on abridged original texts. They had to be evaluated relative to several statements (see Figure 5) representing different aspects of critical engagement with research-based assumptions in Educational Science, such as conclusions from different research approaches (qualitative vs. quatitative research methods, see item a), interpretation of the relationship between variables (item b), or generalizability of findings (item c).

Figure 5: Test item for the competence facet Evidence-based Reasoning. The correct solutions are checked

You are reading the following research abstracts:

A: In a scientific study, N = 100 parents and N = 100 teachers were asked, whether and how school problems in adolescents and difficult family situations are correlated. Standardized questionnaires were used in the anonymous survey study. It could be shown that school problems occur frequently for adolescents with family conflicts.

B: Last year, a student showed increasing problems at school. The adolescent himself, his parents, teachers and two friends were hereto interviewed. With each person, a one-hour interview was conducted. It turned out that the boy suffers from low self-esteem and is shunned by his classmates. This implies that teachers should always take social conditions into account in appraising students.

Please mark, which attributes are rather appropriate for A or B:

	…rather applies to A	…applies to both A and B	…rather applies to B
a) The results give information about the problem in an individual case.	☐	☐	☒
b) A general correlation between the attributes "family conflicts" and "school problems" can be deduced.	☒	☐	☐
c) The results can be generalized to other adolescents.	☒	☐	☐

Due to assumed multidimensionality and aspired scale reliability, an incomplete block design was used in study 1 to minimize strain for participants (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; Shoemaker, 1973). Eight tasks with independent item stems were selected equally from the competence facets (IL, SL, ER) and assigned to one of 20 blocks. For example, the first item block (testlet 1) consisted of one IL item, four SL items, and three ER items, and occurred in booklet 1, 18, 19, and 20. While avoiding redundancy and local dependency, four item blocks were combined in each test booklet. Moreover, half of every booklet was implemented in block inverse order to minimize position effects (J.-T. Kuhn & Kiefer, 2013; see Figure 6).

Figure 6: Booklet design in study 1

booklet	1		2		3		4		…	16		17		18		19		20	
	F	R	F	R	F	R	F	R		F	R	F	R	F	R	F	R	F	R
block position																			
1	1	4	2	5	3	6	4	7		16	19	17	20	1	20	1	19	1	2
2	2	3	3	4	4	5	5	6		17	18	18	19	18	1	2	20	2	1
3	3	2	4	3	5	4	6	5		18	17	19	18	19	18	19	1	3	20
4	4	1	5	2	6	3	7	4		19	16	20	17	20	19	20	2	20	3

Note. 20 booklets were composed of 20 testlets (item blocks), with four-time occurrence of one testlet and four testlet positions per booklet. Each cell represents one testlet with 8 tasks (single items & item sets); the number indicates the position of the block in the test booklet.

Abbreviation: F = forward block order, R = reverse block order.

During test implementation 40 minutes were allotted to complete the test. Ensuring that enough tasks were available, the booklets were composed for an estimated maximum processing time of 60 minutes. To prevent frustration, the test administrators instructed the participants that while it was important to process as many tasks as possible in the given time, it was impossible to complete the full test. Consequently, on average, about 30 % of the tasks were left out, whereof 20 % were not reached, and 10 % were omitted before break-off. In the remaining test time, participants were asked to provide personal and professional background information. Furthermore, potential predictors (e.g., cognitive abilities, self-perceived research ability, motivation) were assessed. For reasons of brevity, the latter results were not addressed in the analyses presented.

In study 2, only one booklet (forward/reverse order, cf. item numbers in Figure 1 to Figure 3) was used. To secure the lecturers' willingness to allow for the data collection again, the test booklets were considerably shorter, with an estimated processing time of 30 minutes. Most of the test items were selected from the item pool of study 1 (see Table 2), with efforts to avoid local item dependency, choose a representative set of requirements in all three ERL facets, and cover as wide a range of item difficulty as possible. To broaden the scope of the competence facets (e.g., IL: Formulation of research questions), three additional tasks were included, which originally were not considered in study 1 due to lack of space or incomplete revision.

In the current analyses, some items were excluded due to poor item fit (0.80 ≥ Infit/Outfit ≥ 1.20, cf. Adams & Wu, 2002). Thus, the results from study 1 were based on 226 test items with reference to 141 item stems (*unadjusted)* or respectively, after exclusion of poor fitting items, on 193 items with reference to 119 stems (*adjusted*). In study 2, the unadjusted data set included 29 test items with reference to 18 stems, while the adjusted data set contained 22 items with reference to 14 item stems. The distribution of test items to the three competence facets IL, SL and ER can be found in Table 2.

Table 2: Distribution of the test items to the competence facets Information Literacy, Statistical Literacy, and Evidence-based Reasoning before and after exclusion of poor fitting items in the standardization study and the study in summer semester 2014.

	Winter semester 2012/2013 and summer semester 2013 (study 1)		Summer semester 2014 (study 2)	
	unadjusted (n_i = 226)	adjusted (n_i = 193)	unadjusted (n_i = 29)	adjusted (n_i = 22)
Competence facets				
IL	32 (14.2%)	30 (15.5%)	7 (24.1%)	6 (27.3%)
SL	85 (37.6%)	71 (36.8%)	8 (27.6%)	6 (27.3%)
ER	109 (48.2%)	92 (47.4%)	14 (48.3%)	10 (45.5%)

Note. IL = Information Literacy; SL = Statistical Literacy; ER = Evidence-based Reasoning; n_i = number of test items included.

4.3 Statistical analyses

In competency assessment, psychometric models usually are utilized to analyze factorial models (Hartig & Höhler, 2009; Wilson, 2005). Popular psychometric models based on modern test theory (as opposed to classical test theory) include item response theory (IRT), which rest upon stringent statistical assumptions (i.e., monotonicity, local independence, unidimensionality). *Monotonicity* asserts that the likelihood of successful performance is a non-decreasing function of a test taker's proficiency. *Local independence* infers that item performance is provisionally independent given an examinee's trait level, whereas the *dimensionality* of an assessment refers to the quantity of latent aptitudes required to capture the construct of interest (Embretson & Reise, 2000). Multidimensional IRT models (Hartig & Höhler, 2009; Wei, 2008) assume several latent dimensions that are represented – in case of between-item dimensionality (Hartig & Höhler, 2008) – by item clusters, which in turn also can be treated as unidimensional sub-constructs. Accordingly, in model 2 (see section 3), each test item was assigned to only one of the three proposed competence facets: IL, SL, or ER.

It has been postulated (e.g., Gustafson, 2001; Humphreys, 1985) that the assumption of strict unidimensionality is not applicable, for example, to educational and psychological assessment where, in addition to one dominant latent trait, other minor latent factors likely influence participants' responses. To separate dominant dimensions from transient dimensions, the concept of *essential unidimensionality* was proposed by Stout (1987). Essential unidimensionality can be conceptualized as the least complex test structure necessary to allow for the assumptions of monotonicity and local independence to be met, and thereby relaxing some of the stringent assumptions of IRT models. Corresponding models can be implemented by so-called bifactor models (Holzinger & Swineford, 1937), which allow each item response to be explained by both a dominant factor and secondary orthogonal fac-

tors (Gibbons & Hedeker, 1992). The dominant trait is the factor of interest (i.e., ERL), whereas the secondary traits (i.e., IL, SL, ER) may be considered as subdomains. In model 3 and model 4 (see section 3), each item loads on the general factor and only one of the subdomain factors. Moreover, the subdomains are orthogonal to each other and to the dominant factor. The underlying assumption of such "restricted" bifactor models (Reise et al., 2010) is that all items measure a common latent trait, such as ERL, but that the variance of each item also is influenced by additional common factors caused by "parcels" of items drawing from similar aspects of the underlying traits. For this reason, items that were included in more than one testlet were excluded from the comparison of model 3 and model 4.

To identify the best fitting model, the four competing models (see section 3) were analyzed with the R package Test Analysis Modules (TAM; Kiefer, Robitzsch, & Wu, 2016). For model selection, the information criteria *Akaike Information Criterion* (AIC; cf. Akaike, 1974, 1987), *Bayesian Information Criterion* (BIC; e.g., Read & Cressie, 1988; Wasserman, 2000) and *Consistent Akaike Information Criterion* (CAIC; e.g., Bozdogan, 1987) were used, with the latter particularly recommended as robust estimator. As a decision rule, the model with the lowest values was the best fit to the data (e.g., Schermelleh-Engel, Moosbrugger, & Müller, 2003). By default, TAM treats missing values as ignorable. But in study 1, a two-stage-procedure to handle omitted and non-reached responses was implemented (e.g., PISA: Adams & Wu, 2002; TIMSS: Martin, Gregory, & Stemler, 2000; cf. Köhler, Pohl, & Carstensen, 2014): For item calibration (study 1), all missing values were treated as ignorable (condition a). With advancing analysis and the need for estimating person ability parameters, omitted responses were scored as incorrect, whereas not-reached responses were left as missing (condition b). The rationale for handling missing data in *condition b* was that omitted responses occur when participants unintentionally skip a task or decide consciously against answering it (Ludlow & O'Leary, 1999), and that in such cases a random answer would most probably result in an incorrect answer (Educational Testing Services, 2014). In psychometrics, it is common for reliability to be estimated by coefficient α, KR-20, or Spearman-Brown corrected split-half correlations. The precision of person estimates in the current study was reported by the EAP/PV (expected a posteriori/ plausible value) reliability coefficient which represents the explained variance in the estimated model divided by total person variance, and is comparable with Cronbach's α (Bond & Fox, 2006; J. Rost, 2004; Walter, 2005). Reliability coefficients of .75 or higher are considered good, although values of at least .55 are deemed satisfactory for group comparisons (Rost, 2013). For multidimensional constructs, however, determination of the alpha coefficient is complex, thus alternate indices need to be applied. Omega (ω) is a model based reliability estimate that combines higher-order and lower-order factors. Though in the case of a bifactor model, it is necessary to separately estimate the reliability of the broad general dimension as well as the specific group dimensions with the influences of the others removed. Omega-hierarchical (ωh) is the model based reliability estimate of one target construct with others removed. The value of omega and/or omega-

hierarchical may assist in determining which composite scales possess sufficient reliable variance to be interpreted; therefore, Green and Yang (2009, 2015) recommended reporting both coefficients.

5. Results

The main purpose of this study was to examine whether ERL can be modeled as one-dimensional latent construct (model 1); as multidimensional ability which is composed of the three subdimensions IL, SL, and ER (model 2); or as multidimensional ability with one dominant factor G and secondary factors, which represent either the competence facets IL, SL and ER (model 3) or the testlet structure (model 4).

The results of the comparison of model 1 to model 3 are outlined in Table 3. In study 1 and study 2, the data were initially analyzed based on the full item set (*unadjusted*) and, after exclusion of misfitting items, on a reduced item set (*adjusted*). In addition, the unadjusted and adjusted item sets were analyzed both under condition a (i.e., omitted and not-reached items were treated as ignorable) and condition b (i.e., omitted items were scored as incorrect responses). The values of the information criteria indicate a similar trend across study 1 and study 2. Compared to the one- and three-dimensional IRT models, the bifactor model solution in model 3 appears to be better fitting because the corresponding values of AIC, BIC and CAIC were lowest. In most cases, the information criteria values of the three-dimensional model were closer to the superior bifactor model than to the one-dimensional model. The only exception was the model comparison for study 1 under condition a (both adjusted and unadjusted), where the information criteria values of the one-dimensional and the bifactor model were closer to each other than to the three-dimensional model.

In the comparison of the two bifactor models (see Table 4), model 4 was superior to model 3 under condition b. This indicates that the testlets were perceived as differently motivating (Eklöf, 2010; Marentette, Meyers, Hurtz, & Kuang, 2012). Further, recoding omitted responses as incorrect appeared to cause a statistical artifact when modeling the factor structure, at least in this sample as this testlet effect was not inherent in the data originally. Collectively, the reported results favor model 3. Item intercepts and standardized factor loadings of the corresponding bifactor solution of the adjusted data sets from study 1 and 2 are displayed in Table 5 along with the different reliability coefficients. As expected, the intercepts were higher when omitted responses were scored as incorrect under condition b compared to condition a, indicating that the test items were scaled as more difficult on the ability continuum. In contrast, the test booklet in study 2 contained relatively more difficult IL and SL items; however, probably due to the higher proportion of easier items for the competence facet ER (see Table 2), this test booklet turned out easier than the booklet in study 1.

Table 3: Goodness-of-fit statistics for comparing competing models of the test instrument in study 1 and study 2

Sample	Model	Factors	Final Deviance	n_p	AIC	BIC	CAIC
Study 1, condition a							
unadjusted [a] (n_i = 226)	1	1 (G)	47590.2	227	48044	49228	49455
	2	3 (IL, SL, ER)	47592.0	232	48056	49266	49498
	3	4 (G, IL, SL, ER)	47561.2	230	**48021**	**49221**	**49451**
adjusted [b] (n_i = 193)	1	1 (G)	43049.0	194	43437	44449	44643
	2	3 (IL, SL, ER)	43052.4	199	43450	44488	44687
	3	4 (G, IL, SL, ER)	43020.1	197	**43414**	**44442**	**44639**
Study 1, condition b							
unadjusted [a] (n_i = 226)	1	1 (G)	56564.1	227	57018	58202	58429
	2	3 (IL, SL, ER)	56425.4	232	56889	58099	58331
	3	4 (G, IL, SL, ER)	56402.0	230	**56862**	**58062**	**58292**
adjusted [b] (n_i = 193)	1	1 (G)	51441.6	194	51830	52841	53035
	2	3 (IL, SL, ER)	51304.5	199	51703	52740	52939
	3	4 (G, IL, SL, ER)	51280.8	197	**51675**	**52702**	**52899**
Study 2, condition a							
unadjusted [a] (n_i = 27)	1	1 (G)	18439.2	28	18495	18625	18653
	2	3 (IL, SL, ER)	18383.8	33	18450	18602	18635
	3	4 (G, IL, SL, ER)	18367.6	31	**18430**	**18573**	**18604**
adjusted [b] (n_i = 22)	1	1 (G)	15871.4	23	15917	16024	16047
	2	3 (IL, SL, ER)	15830.4	28	15886	16016	16044
	3	4 (G, IL, SL, ER)	15816.6	26	**15869**	**15989**	**16015**
Study 2, condition b							
unadjusted [a] (n_i = 27)	1	1 (G)	19917.8	28	19974	20103	20131
	2	3 (IL, SL, ER)	19807.7	33	19874	20026	20059
	3	4 (G, IL, SL, ER)	19796.0	31	**19858**	**20001**	**20032**
adjusted [b] (n_i = 22)	1	1 (G)	17171.6	23	17218	17324	17347
	2	3 (IL, SL, ER)	17097.0	28	17153	17283	173101
	3	4 (G, IL, SL, ER)	17087.4	26	**17139**	**17260**	**17286**

Note. Study 1: winter semester 2012/2013 and summer semester 2013. Study 2: summer semester 2014. *N* (study 1) = 1360; *N* (study 2) = 753. Under condition a, both omitted and not-reached items were treated as ignorable. Under condition b, omitted items were recoded as incorrect response and not-reached items left as missing data.

n_i = number of test items included; n_p = number of estimated parameters; G = general factor Educational Research Literacy; IL = Information Literacy; SL = Statistical Literacy; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion. The parameters of the respective best fitting solution are indicated in bold.

[a] Unadjusted: All test items are included in the data set.

[b] Adjusted: Only items with good model fit (0.80 ≤ Infit/Outfit ≤ 1.20; cf. Adams & Wu, 2002) are included.

Table 4: Goodness-of-fit statistics for comparing competing models of the test instrument in study 1 and study 2

Sample	Model	Factors	Final Deviance	n_p	AIC	BIC	CAIC
Study 1, condition a							
unadjusted [a] (n_i = 208)	3	4 (G, IL, SL, ER)	40393.22	212	**40817**	**41922**	**42134**
	4	21 (G, testlet 1–20)	40381.86	229	40840	42033	42262
adjusted [b] (n_i = 177)	3	4 (G, IL, SL, ER)	36525.83	181	**36888**	**37831**	**38012**
	4	21 (G, testlet 1–20)	36528.33	198	36924	37956	38154
Study 1, condition b							
unadjusted [a] (n_i = 208)	3	4 (G, IL, SL, ER)	47392.77	212	47817	48921	49133
	4	21 (G, testlet 1–20)	47167.41	229	**47625**	**48819**	**49048**
adjusted [b] (n_i = 177)	3	4 (G, IL, SL, ER)	43046.37	181	43408	44352	44533
	4	21 (G, testlet 1–20)	42854.89	198	**43251**	**44283**	**44481**

Note. Study 1: winter semester 2012/2013 and summer semester 2013. *N* = 1360. A booklet design was used, in which 20 testlets (item blocks) occurred on one of four possible positions in different booklets (see Figure 6). Tasks that occurred in more than one testlet had to be excluded from analysis because of violating the assumption of the restricted bifactor models (see section 4.3). Under condition a, both omitted and not-reached items were treated as ignorable. Under condition b, omitted items were recoded as incorrect response and not-reached items left as missing data.

n_i = number of test items included; n_p = number of estimated parameters; G = general factor Educational Research Literacy; IL = Information Literacy; SL = Statistical Literacy; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; CAIC = Consistent Akaike Information Criterion. The parameters of the respective best fitting solution are indicated in bold.

[a] Unadjusted: All test items are included in the data set.

[b] Adjusted: Only items with good model fit (0.80 ≤ Infit/Outfit ≤ 1.20; cf. Adams & Wu, 2002) are included.

The standardized factor loadings on the general factor proved to be of medium size and consistently were smaller for the subdimensions IL, SL and ER (see Table 5). The comparison of the different treatments of omitted responses indicated that *condition b* led to a higher item-scale correlation, especially for the subdimension ER. This was further supported by the improved EAP-reliability for this factor (condition a: EAP-reliability = .07; condition b: EAP-reliability = .31); however, the EAP-reliability of all three subdimensions (i.e., IL, SL, ER) was low compared to prevailing standards. In contrast, reliability of the general factor was found to be satisfactory as demonstrated by good reliability for study 1 and as satisfactory reliability for study 2. According to the omega hierarchical coefficient almost 90 % in study 1 and approximately 70 % in study 2 of the variance in raw scale scores could be explained by the variation in the general factor in study 1 and in study 2, respectively. The lower omega hierarchical found in study 2 may be due to a smaller sample size as compared to study 1.

Table 5: Standardized factor loadings and reliability for the four-dimensional bifactor solution (model 3) of the adjusted data sets from study 1 and 2

		Intercepts *M* (*SD*)	Standardized factor loadings				EAP-reliability				ω	ωh
			G	IL	SL	ER	G	IL	SL	ER		
Study 1a	IL	0.01 (1.31)	.30	.15								
	SL	-0.47 (1.48)	.30		.20							
	ER	-0.03 (0.99)	.31			.12						
	total	-0.19 (1.25)					.56	.05	.16	.07	.92	.87
Study 1b	IL	0.40 (1.32)	.36	.14								
	SL	-0.20 (1.43)	.36		.22							
	ER	0.49 (0.98)	.35			.27						
	total	0.22 (1.25)					.63	.05	.18	.31	.95	.82
Study 2a	IL	0.92 (0.86)	.35	.24								
	SL	0.08 (1.03)	.34		.18							
	ER	-0.81 (0.59)	.34			.29						
	total	-0.09 (1.18)					.51	.13	.08	.25	.67	.66
Study 2b	IL	1.10 (0.83)	.36	.24								
	SL	0.36 (0.89)	.36		.27							
	ER	-0.68 (0.60)	.35			.31						
	total	0.09 (1.17)					.53	.13	.17	.29	.70	.66

Note. Study 1: winter semester 2012/2013 and summer semester 2013, $N = 1360$, $n_i = 193$. Study 2: summer semester 2014, $N = 753$, $n_i = 22$. Under condition a, both omitted and not-reached items were treated as ignorable. Under condition b, omitted items were recoded as incorrect response and not-reached items as ignorable.

G = general factor Educational Research Literacy; IL = Information Literacy; SL = Statistical Literacy; ER = Evidence-based Reasoning; EAP/PV reliability = expected a posteriori/plausible value reliability; ω = reliability coefficient Omega; ωh = reliability coefficient Omega hierarchical.

6. Conclusions

Three analysis strategies were employed to investigate the assumed dimensionality of ERL. The most appropriate test structure was identified by comparing different competing competence structure models. The generalizability of the results was ascertained by comparing the factorial structure in two independent samples. Lastly, the impact of different treatment methods for omitted responses was examined in different models.

The analysis of competing one- and multidimensional competence models revealed the four-dimensional bifactor model was superior to the other models in explaining data structure in two independent samples (question 1 and 2). Accordingly, essential multidimensionality of Educational Research Literacy (ERL) is to be assumed. The bifactor model could serve an acceptable compromise be-

tween the unidimensionality preference and the multidimensionality reality. An appealing feature of the bifactor model is that it allows for simultaneous evaluation of both the general and specific influences on indicators (subdimensions). The bifactor results showed that each dimension of ERL is confounded by both general and specific sources of variance (model 3), indicating that this ability seems to consist of one general factor and the secondary factors of Information Literacy (IL), Statistical Literacy (SL) and Evidence-based Reasoning (ER). The dominant factor represents the generic aspect of ERL in relation to the research cycle, and presumably comprises something like reflection ability (Körkkö, Kyrö-Ämmälä, & Turunen, 2016). Jay and Johnson (2002) described reflection ability as three steps: descriptive, comparative, and critical reflection. In contrast, the subdomain factors represent particular requirements of the different research steps: Information search, which usually is guided by a certain research question, demands different abilities (e.g., identification of semantically relevant keywords) than engagement with statistical/numerical information (e.g., in the form of tables). Critically evaluating evidence-based assumptions eventually necessitates the application of research-methodological background knowledge.

The results of the current analyses also explain the contradictory findings supporting both a one-dimensional model of ERL (Groß Ophoff et al., 2014) and a three-dimensional Rasch model with the (highly intercorrelated) subdimensions IL, SL, and ER (Groß Ophoff et al., 2015). The different model fit patterns depending on the treatment of omitted responses (question 3) emphasize the importance of accounting for the influence of missing data (Custer, Sharairi, & Swift, 2012; Köhler et al., 2014; Rose, von Davier, & Xu, 2010). In the study 1 sample, the number of omitted responses correlated only to a negligible effect of $r = .15$ ($p < .001$) with the weighted likelihood estimates (WLE) of ability in the one-dimensional solution under condition a (both omitted and not-reached items were treated as ignorable). After scoring omitted responses as incorrect (condition b), the correlation increased to $r = -.60$ ($p < .001$), implying that the parameter estimates underestimate the actual ability. It might be argued that under this condition the internal consistency of the test (EAP-reliability, see Table 4) could be slightly improved; however, it has been shown (Chang & Wang, 2010; Eckes, 2015; Wainer & Wang, 2000) that neglecting testlet effects (which is the case for model 3 under condition b) may not lead only to underestimated standard errors of ability parameters and biased estimates of both item discrimination and item difficulty, but also produce a higher measurement accuracy. Collectively, the findings support the recommendations of Rose et al. (2010) that omissions should not be scored as incorrect.

Given that the general factor of ERL is dominant over the secondary factors, essential unidimensionality can be assumed (Stout, 1987). Thus a one-dimensional model can be applied, for example, for the assessment and feedback about learning gains on student level (Hartig, 2008; Wilson & Scalise, 2006), but without further differentiation of the three subdimensions because of their low reliability. This can be explained by the fact that the general factor is implicitly partialled out (cf. Li, Jiao, & Lissitz, 2014), so that only the remaining variance can be used to calculate

the reliability of the secondary factors in the bifactor model. However, this diagnostic issue is only of minor importance, since the question of within-item multidimensionality focuses on the analytical investigation of separable information components.

In large-scale assessments, one-dimensional competence models have been proved as adequate and substantial description of the data structure, for example in the field of research on Statistical Literacy (e.g., Watson & Callingham, 2003). Compared to multidimensional competence models, one-dimensional models facilitate criterion-referenced interpretation and are more easily conveyed to educational practice (e.g., Groß Ophoff, Isaac, Hosenfeld, & Eichler, 2008). If the general factor was not as dominant, then this would warrant analyzing the subdimensions separately; however, it might be conceivable to use the three-dimensional model (model 2) as basis for analysis with the objective of course or study program development. The reliability coefficients for the subdimensions in this model are higher (e.g., EAP-reliability of the adjusted solution, condition a: IL = .40; SL = .54; ER = .53) due to the underlying G-factor. The reliability could be improved further by including high discriminating items in assembling test booklets for future studies. Moreover, it seems worthwhile to analyze the specific task requirements in-depth with reference to the appointed subscales, and to develop the test further based on this analysis (cf. Schladitz, Groß Ophoff, & Wirtz, in this issue).

In the presented analyses, the construct validation of the factorial structure of Educational Research Literacy was paramount; however, generalizability is restricted due to non-probabilistic opportunity samples in study 1 and study 2. This is a typical problem in this research field because students in higher education institutions are difficult to access (Zlatkin-Troitschanskaia, Pant, Kuhn, Toepper, & Lautenbach, 2016). The strength of the presented research lies in the large samples and in the inclusion of several universities from various German federal states. To further advance evidence about ERL, data from the second assessment time point in study 2 and another study at Austrian Universities for Education are being analyzed currently. For future studies, consideration should be given to alternative approaches (e.g., obligation to participate as part of study program development, voluntary participation in a panel study with incentives).

With respect to the conceptual framework, the reported results from the current investigation correspond with the notion of ERL from the perspective of Research-Based Learning (Lambert, 2009), Inquiry-Based Learning (Pedaste et al., 2015), or Problem-Based Learning (Hmelo-Silver, 2004), where learning emerges from a holistic research process. Accordingly, learning gains in ERL cannot be characterized solely by specific competence facets (c.f. model 2), but by progressive change of perspective from action-oriented coping with everyday practice to a more academic evidence-oriented attitude. It should be noted, therefore, that focusing on imparting only particular competence facets such as basic skills in Statistical Literacy, is too narrow of a view relative to the objectives of Higher Education (cf. Olehnovica et al., 2015). Instead, the entire competence spectrum should be pursued and learning opportunities should be offered by purposefully increasing study

requirements (e.g., zone of proximal development, Vygotsky, 1978). This is reinforced further by the fact that less proficient students are capable of finding and reproducing research information in tables, diagrams and summaries whereas only advanced students are proficient in evaluating scientific evidence and critically appraising research-related conclusions (Brown, Furtak, et al., 2010; Groß Ophoff et al., 2014; Zeuch et al., 2017). To increase research competencies, it is also critical to foster research self-efficacy and to offer practice-relevant, and therefore meaningful research opportunities, for example, in form of active-participant learning opportunities during studies (e.g., Bard, Bieschke, Herbert, & Eberz, 2000; Bell, 2016; Butcher & Maunder, 2014). Future research should examine combining instructional approaches with continual assessment of student performance (Wilson & Scalise, 2006). Further, more complex and authentic educational settings should be considered for further development of the presented test instrument (e.g., problem-oriented performance tasks, cf. Klein et al., 2007; Wenglein, Baur, Heininger, & Prenzel, 2015). The final step of making logical decisions, taking a position by integrating various evidence (Problem Solving, e.g., Phye, 2001), or transferring research-based insights (Billing, 2007) is not covered by the test instrument presented in this paper.

Following Blömeke et al. (2015), the assessment approach presented in this paper can be referred to as assessment of situation-specific abilities in the field of Educational Science. It is based on the assumption that Higher Education in general and degree programs, particularly in Educational Science, enable students to *engage with research* (Borg, 2010), which establishes options for action in future practice. Whether the dominant factor of ERL is the domain-specific learning outcome of study programs examined here (Eisenhart & DeHaan, 2005; Love, 2009) or a generic, and therefore transdisciplinary, ability (Clanchy & Ballard, 1995; Gilbert, Balatti, Turner, & Whitehouse, 2004) cannot be conclusively determined based on the present results. To resolve this issue, Wecker et al. (2014) proposed three research paradigms: a) expert studies in a subject area outside of the respective expertise, b) experimental transfer studies, and c) correlational studies on predictors of performance controlling for competing explanatory factors. In order to enhance theoretical and empirical-conceptual foundations of Educational Research Literacy, practice-oriented intervention studies seem to be essential, too (e.g., based on Action Research, cf. Altrichter, Feldman, Posch, & Somekh, 2013).

Acknowledgments

This work has been developed in the project Learning the Science of Education (LeScEd, FKZ: 01PK11009A), which was funded by the German ministry of education and research (BMBF) within the research programme KoKoHs (2011-2015). We thank Dr. Alexander Robitzsch for his helplful support concerning questions about the R-package Test Analysis Modules (TAM).

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

Adams, R. J., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: Organisation for Economic Cooperation and Development.

Adedokun, O. A., Bessenbacher, A. B., Parker, L. C., Kirkham, L. L., & Burgess, W. D. (2013). Research skills and STEM undergraduate research students' aspirations for research careers: Mediating effects of research self-efficacy. *Journal of Research in Science teaching, 50*(8), 940–951.

Akaike, H. (1974). A new look at statistical mode identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317–332. doi:10.1007/bf02294359

Altrichter, H., Feldman, A., Posch, P., & Somekh, B. (2013). *Teachers investigate their work: An introduction to action research across the professions*. New York, NY: Routledge.

Barchfeld, P., & Sodian, B. (2009). Differentiating theories from evidence: The development of argument evaluation abilities in adolescence and early adulthood. *Informal Logic, 29*(4), 396–416.

Bard, C. C., Bieschke, K. J., Herbert, J. T., & Eberz, A. B. (2000). Predicting research interest among rehabilitation counseling students and faculty. *Rehabilitation Counseling Bulletin, 44*(1), 48–55.

Bell, R. (2016). The continuing search to find a more effective and less intimidating way to teach research methods in higher education. *Innovations in Education and Teaching International, 53*(3), 285–295.

Benson, A., & Blackman, D. (2003). Can research methods ever be interesting? *Active learning in Higher Education, 4*(1), 39–55. doi:10.1177/1469787403004001004

Ben-Zvi, D., & Garfield, B. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. New York, NY: Kluwer Academic Publishers.

Billing, D. (2007). Teaching for transfer of core/key skills in higher education: Cognitive skills. *Higher Education, 53*(4), 483–516. doi:10.1007/s10734-005-5628-5

Blixrud, J. C. (2003). Project SAILS: Standardized assessment of information literacy skills. *ARL Bimonthly Report*. Retrieved from http://old.arl.org/bm~doc/arl-br230231.pdf

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

Blömeke, S., & Zlatkin-Troitschanskaia, O. (2013). *Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs (KoKoHs Working Papers, 1.)* [Modeling and measuring competencies in higher education: Objectives, theoretical framework, design, and challenges of the BMBF research program KoKoHs]. Retrieved from http://www.kompetenzen-im-hochschulsektor.de/Dateien/KoKoHs_WP1_Bloemeke_Zlatkin-Troitschanskaia_2013_.pdf

Bond, T. G., & Fox, C. M. (2006). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Borg, S. (2010). Language teacher research engagement. *Language Teaching, 43*(4), 391–429.

Borg, S., & Alshumaimeri, Y. (2012). University teacher educators' research engagement: Perspectives from Saudi Arabia. *Teaching and Teacher Education, 28*(3), 347–356.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. doi:10.1007/bf02294361

Braun, E., Gusy, B., Leidner, B., & Hannover, B. (2008). Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp) [The Berlin Evaluation Instrument for self-evaluated student competences]. *Diagnostica, 54*(1), 30–42.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3–12. doi:10.1111/j.1745-3992.2010.00195.x

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education, 11*(3), 301–318.

Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010). The evidence-based reasoning framework: Assessing scientific reasoning. *Educational Assessment, 15*(3/4), 123–141. doi:10.1080/10627197.2010.530551

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*(3/4), 142–174. doi:10.1080/10627197.2010.530562

Butcher, J., & Maunder, R. (2014). Going URB@ N: Exploring the impact of undergraduate students as pedagogic researchers. *Innovations in Education and Teaching International, 51*(2), 142–152.

Calzada Prado, J., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri, 63*(2), 123–134.

Catts, R. & Lau, J. (2008). *Towards information literacy indicators: Conceptual framework paper*. Paris, France: Information Society Division, Communication and Information Sector, UNESCO.

Chang, Y., & Wang, J. (2010, July). *Examining testlet effects on the PIRLS 2006 assessment*. Paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden.

Clanchy, J., & Ballard, B. (1995). Generic skills in the context of higher education. *Higher Education Research and Development, 14*(2), 155–166.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*(4), 609–624. doi:http://dx.doi.org/10.1016/j.cedpsych.2007.10.002

Colquitt, J. A., & Zapata-Phelan, C. P. (2007). Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal. *Academy of Management Journal, 50*(6), 1281–1303. doi:10.5465/amj.2007.28165855

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Custer, M., Sharairi, S., & Swift, D. (2012). *A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilizing the Rasch model and joint maximum likelihood estimation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, British Columbia.

Davidson, K. (2013). Teachers' reported utilization of reading disabilities research. *Alberta Journal of Educational Research, 59*(3), 487–502.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47*(2), 108–121.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272. doi:10.1007/s11092-015-9233-6

Dunn, D. S., Halonen, J. S., & Smith, R. A. (Eds.). (2008). *Teaching critical thinking in psychology: A handbook of best practices*. Malden, MA: Wiley-Blackwell.

Durrant, G. B. (2005). *Imputation methods for handling item – nonresponse in the social sciences: A methodological review*. NCRM Methods Review Papers. NCRM/002. Southhampton, England: National Centre for Research Methods and

Statistical Sciences Research Institute (S3RI). Retrieved from http://eprints.ncrm.ac.uk/86/

Eckes, T. (2015). Lokale Abhängigkeit von Items im TestDaF-Leseverstehen: Eine Testlet-Response-Analyse [Local item dependence in the TestDaF reading section: A testlet response analysis]. *Diagnostica, 61*(2), 93–106. doi:10.1026/0012-1924/a000118

Educational Testing Services (2014). *A guide to understanding the literacy assessment of the STEP Skills Measurement Survey*. Princeton, NJ: IEA-ETS Research Institute.

Eisenhart, M., & DeHaan, R. L. (2005). Doctoral preparation of scientifically based education researchers. *Educational Researcher, 34*(4), 3–13.

Eklöf, H. (2010). *Student motivation and effort in the Swedish TIMSS advanced field study*. Paper presented at the 4th meeting of the IEA International Research Conference, Gothenburg, Sweden.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Federal Ministry of Education and Research. (2015). Bericht der Bundesregierung über die Umsetzung des Bologna-Prozesses 2012–2015 in Deutschland [Federal Government's report on the implementation of the Bologna process 2012–2015 in Germany]. Retrieved from https://www.bmbf.de/files/Bericht_der_Bundesregierung_zur_Umsetzung_des_Bologna-Prozesses_2012-2015.pdf

Frey, A., Hartig, J., & Rupp, A. 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.

German Science Council. (2000). Empfehlungen zur Einführung neuer Studienstrukturen und -abschlüsse (Bakkalaureus/Bachelor – Magister/Master) in Deutschland [Recommendations about the introduction of new study structures and degrees (Bachelor/Master) in Germany]. Retrieved from http://www.wissenschaftsrat.de/download/archiv/4418-00.pdf

Gess, C., Wessels, I., & Blömeke, S. (2017). Domain-specificity of research competencies in the social sciences: Evidence from differential item functioning. *Journal for Educational Research Online*, *9*(2).

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436.

Gilbert, R., Balatti, J., Turner, P., & Whitehouse, H. (2004). The generic skills debate in research higher degrees. *Higher Education Research and Development, 23*(3), 375–388.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In M. von Davier & E. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 3, pp. 125–156). Hamburg, Germany: IEA-ETS Research Institute.

Gotch, C. M., & French, B. F. (2013). Elementary teachers' knowledge and self-efficacy for measurement concepts. *The Teacher Educator, 48*(1), 46–57.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice, 33*(2), 14–18. doi:10.1111/emip.12030

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167. doi:10.1007/s11336-008-9099-3

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice, 34*(4), 14–20. doi:10.1111/emip.12100

Groß Ophoff, J., Isaac, K., Hosenfeld, I., & Eichler, W. (2008). Erfassung von Leseverständnis im Projekt VERA [Assessment of reading literacy in mandato-

ry performance test in German primary education]. In B. Hofmann & R. Valtin (Eds.), *Checkpoint literacy. Tagungsband 2 zum 15. Europäischen Lesekongress 2007 in Berlin [Proceedings of the 15th European conference on reading]* (pp. 36–51). Berlin, Germany: Deutsche Gesellschaft für Lesen und Schreiben (DGLS).

Groß Ophoff, J., Schladitz, S., Leuders, J., Leuders, T., & Wirtz, M. (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.

Groß Ophoff, J., Schladitz, S., Lohrmann, K., & Wirtz, M. (2014). Evidenzorientierung in bildungswissenschaftlichen Studiengängen: Entwicklung eines Strukturmodells zur Forschungskompetenz [Evidence-orientation in educational science degree programs: Development of a structure model of educational research literacy]. In W. Bos, K. Drossel, & R. Strietholt (Eds.), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen* [Empirical educational research and evidence-based reforms in eduation] (pp. 251–276). Münster, Germany: Waxmann.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38*(5), 427–437.

Grundmann, R., & Stehr, N. (2012). *The power of scientific knowledge: From research to Public Policy*. New York, NY: Cambridge University Press.

Gustafson, J.-E. (2001). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 77–101). New York, NY: Routledge.

Hammersley, M. (2004). Some questions about evidence-based practice in education. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education* (pp. 133–149). Maidenhead, England: Open University Press.

Harper, D. J., Gannon, K. N., & Robinson, M. (2012). Beyond evidence-based practice: Rethinking the relationship between research, theory and practice. In R. Bayne & G. Jinks (Eds.), *Applied psychology: Research, training and practice* (2nd ed.). London, England: Sage.

Hartig, J. (2008). Psychometric models for the assessment of competencies. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 69–90). Göttingen, Germany: Hogrefe.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie, 21*(6), 89–101.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies In Educational Evaluation, 35*(2–3), 57–63.

Heinze, N. (2008). *Bedarfsanalyse für das Projekt i-literacy: Empirische Untersuchung der Informationskompetenz der Studierenden der Universität Augsburg [*Requirement analysis for the project i-literacy: Empirical investigation of the information literacy of students at the university Augsburg]. Retrieved from http://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/index/index/docId/685

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235–266. doi:10.1023/B:EDPR.0000034022.16470.f3

Holzinger, K., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41–54.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 201–224). New York, NY: Wiley.

Jay, J. K., & Johnson, K. L. (2002). Capturing complexity: A typology of reflective practice for teacher education. *Teaching and Teacher Education, 18*(1), 73–85.

Jindal-Snape, D., Hannah, E., Smith, E., Barrow, W., & Kerr, C. (2009). An innovative practitioner research model of continuing professional development. A case study

of an educational psychologists' professional development programme in Scotland. *School Psychology International, 30*(3), 219–235.
Kiefer, T., Robitzsch, A., & Wu, M. (2016). Test analysis modules (TAM) (Version 1.995-0 (2016-05-31)). Retrieved from http://www.edmeasurementsurveys.com/TAM/Tutorials/
Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment. Facts and fantasies. *Evaluation Review, 31*(5), 415–439.
Köhler, C., Pohl, S., & Carstensen, C. H. (2014). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement, 75*(5), 850–874. doi:10.1177/0013164414561785
Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling, 57*(4), 499–522.
Kopp, B., & Martschinke, S. (2011). Kinder mit deutscher und nicht-deutscher Familiensprache – Ergebnisse aus der CHARLIE-Studie zum Umgang mit migrationsbedingten Disparitäten [Children with German and non-German family language from the CHARLIE-study on handling migration-related disparities]. *Zeitschrift für Grundschulforschung, 4*(2), 46–59.
Körkkö, M., Kyrö-Ämmälä, O., & Turunen, T. (2016). Professional development through reflection in teacher education. *Teaching and Teacher Education, 55*, 198–206.
Kretzschmar, A., Neubert, J. C., Wüstenberg, S., & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence,* 54, 55–69.
Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*(2), 16–46.
Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development, 23*(4), 435–451.
Kuhn, J.-T., & Kiefer, T. (2013). Optimal test assembly in practice. The design of the Austrian educational standards assessment in mathematics. *Zeitschrift für Psychologie, 221*(3), 190–200.
Lambert, C. (2009). Pedagogies of participation in higher education: A case for research-based learning. *Pedagogy, Culture & Society, 17*(3), 295–309. doi:10.1080/14681360903194327
Lawson, T. J. (1999). Assessing psychological critical thinking as a learning outcome for psychology majors. *Teaching of Psychology, 26*(3), 207–209. doi:10.1207/S15328023TOP260311
Lea, M. R., & Street, B. V. (2006). The "academic literacies" model: Theory and applications. *Theory Into Practice, 45*(4), 368–377. doi:10.1207/s15430421tip4504_11
Li, Y., Jiao, H. & Lissitz, R. W. (2014). Applying multidimensional item response theory models in validating test dimensionality: An example of K–12 large-scale science assessment. *Journal of Applied Testing Technology, 13(2)*. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/48367
Love, K. (2009). Literacy pedagogical content knowledge in secondary teacher education: Reflecting on oral language and learning across the disciplines. *Language & Education: An International Journal, 23*(6), 541–560. doi:10.1080/09500780902822942
Lowman, R. L., & Williams, R. E. (1987). Validity of self-ratings of abilities and competencies. *Journal of Vocational Behavior, 31*(1), 1–13. doi:http://dx.doi.org/10.1016/0001-8791(87)90030-3
Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*(4), 615–630. doi:10.1177/0013164499594004

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen [Coping with missing data in psychological research. Problems and solutions]. *Psychologische Rundschau, 58*(2), 103–117.

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education, 60*, 366–376.

Marentette, B. J., Meyers, L. S., Hurtz, G. M., & Kuang, D. C. (2012). Order effects on situational judgment test items: A case of construct-irrelevant difficulty. *International Journal of Selection and Assessment, 20*(3), 319–332. doi:10.1111/j.1468-2389.2012.00603.x

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed., pp. 287–303). Mahwah, NJ: Lawrence Erlbaum.

McMillan, J. H., & Schumacher, S. (2010). *Research in education. Evidence-based inquiry* (7th ed.). Upper Saddle River, NJ: Pearson.

Meltzoff, J. (2010). *Critical thinking about research: Psychology and related fields*. Washington, DC: American Psychological Association.

Ministry of Cultural Affairs of Baden-Württemberg. (2011). *Verordnung des Kultusministeriums über die Erste Staatsprüfung für das Lehramt an Grundschulen (Grundschullehramtsprüfungsordnung I – GPO I)* [Regulation of the first state examination for primary school teachers]. Retrieved from https://www.ph-freiburg.de/fileadmin/dateien/zentral/studienplanung/gpo1_2011.pdf

Naumann, J., Artelt, C., Schneider, W., & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009 [Reading literacy from PISA 2000 to PISA 2009]. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt* [PISA 2009. Taking stock after a decade] (pp. 23–72). Münster, Germany: Waxmann.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. Los Angeles, CA: Cambridge Assessment, Sage.

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science, 12*(2), 123–145.

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. Holyoak & B. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 321–349). New York, NY: Cambridge University Press

Ntuli, E., & Kyei-Blankson, L. (2016). Improving K–12 online learning: Information literacy skills for teacher candidates. *International Journal of Information and Communication Technology Education (IJICTE), 12*(3), 38–50. doi:10.4018/IJICTE.2016070104

O'Connor, L. G., Radcliff, C. J., & Gedeon, J. A. (2002). Applying systems design and item response theory to the problem of measuring information literacy skills. *College & Research Libraries, 63*(6), 528–543. doi:10.5860/crl.63.6.528

Olehnovica, E., Bolgzda, I., & Kravale-Pauliņa, M. (2015). Individual potential of doctoral students: Structure of research competences and self-assessment. *Procedia – Social and Behavioral Sciences, 174*, 3557–3564. doi:http://dx.doi.org/10.1016/j.sbspro.2015.01.1072

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47–61.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525–556.

Phye, G. D. (2001). Problem-solving instruction and problem-solving transfer: The correspondence issue. *Journal of Educational Psychology, 93*(3), 571–578. doi:10.1037/0022-0663.93.3.571

Prenzel, M., Walter, O., & Frey, A. (2007). PISA misst Kompetenzen. Eine Replik auf Rindermann (2006): Was messen internationale Schulleistungsstudien? [PISA measures competencies. A reply to Rindermann (2006): What do international school performance tests measure?] *Psychologische Rundschau, 58*(2), 128–136.

Rankin, J., & Becker, F. (2006). Does reading the research make a difference? A case study of teacher growth in FL German. *The Modern Language Journal, 90*(3), 353–372.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Berlin, Germany: Springer.

Reeves, T. D., & Honig, S. L. (2015). A classroom data literacy intervention for pre-service teachers. *Teaching and Teacher Education, 50*, 90–101.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment, 92*(6), 544–559. doi:10.1080/00223891.2010.496477

Reise, S. P., & Revicki, D. A. (2014). *Handbook of item response theory modeling: Applications to typical performance assessment*: New York, NY: Routledge.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)*. Retrieved from https://www.ets.org/Media/Research/pdf/RR-10-11.pdf

Rost, D. H. (2013). *Interpretation und Bewertung pädagogisch-psychologischer Studien* (3rd completely revised ed.) [Interpretation and evaluation of educational-psychological studies]. Bad Heilbrunn, Germany: Klinkhardt.

Rost, J. (2004). *Lehrbuch Testtheorie/Testkonstruktion* [Textbook test theory/test construction]. (2nd ed.). Bern, Switzerland: Huber.

Rott, B., Leuders, T., & Stahl, E. (2015). Assessment of mathematical competencies and epistemic cognition of pre-service teachers. *Zeitschrift für Psychologie, 223*(1), 39–46.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research – Online, 8*(2), 23–74. Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue20/art2/mpr130_13.pdf

Schladitz, S., Groß Ophoff, J., & Wirtz, M. (2015). Konstruktvalidierung eines Tests zur Messung bildungswissenschaftlicher Forschungskompetenz [Constructive validation of tests for measurement of educational scientific research competence]. *Zeitschrift für Pädagogik, 61* (Supplement), 167–184.

Schladitz, S., Groß Ophoff, J., & Wirtz, M. (2017). Effects of different response formats in measuring Educational Research Literacy. *Journal for Educational Research Online, 9*(2).

Schladitz, S., Rott, B., Winter, A., Wischgoll, A., Groß Ophoff, J., Hosenfeld, I., Wittwer, J. (2013). LeScEd – Learning the science of education. research competence in educational sciences. In S. Blömeke & O. Zlatkin-Troitschanskaja (Eds.), *The German funding initiative "modeling and measuring competencies in higher education": 23 research projects on engineering, economics and social sciences, educa-*

tion and generic skills of higher education students (pp. 82–84). Berlin & Mainz, Germany: Humboldt Universität & Johannes Gutenberg Universität.

Schmid, S., Richter, T., Berthold, K., Bruns, K., & von der Mühlen, S. (2013). KOSWO – students' competencies when dealing with scientific primary literature. In S. Blömeke & O. Zlatkin-Troitschanskaja (Eds.), *The German funding initiative "modeling and measuring competencies in higher education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students* (pp. 71–74). Berlin & Mainz, Germany: Humboldt Universität & Johannes Gutenberg Universität.

Schweizer, K., Steinwascher, M., Moosbrugger, H., & Reiss, S. (2011). The structure of research methodology competency in higher education and the role of teaching teams and course temporal distance. *Learning and Instruction, 21*(1), 68–76.

Shaneyfelt, T., Baum, K. D., Bell, D., Feldstein, D., Houston, T. K., Kaatz, S., Green, M. (2006). Instruments for evaluating education in evidence-based practice. *JAMA: The Journal of the American Medical Association, 296*(9), 1116–1127. doi:10.1001/jama.296.9.1116

Shank, G., & Brown, L. (2007). *Exploring educational research literacy*. New York, NY: Routledge.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger Publishing Company.

Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004* [Standards for teacher education: Educational science]. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. (2005). *Qualifikationsrahmen für Deutsche Hochschulabschlüsse* [Qualifications framework for German university degrees]. Retrieved from http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2005/2005_04_21-Qualifikationsrahmen-HS-Abschluesse.pdf

Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. (2015). *Sachstand in der Lehrerbildung. Stand 21.09.2015* [State of affairs in Teacher Training]. Retrieved from https://www.kmk.org/dokumentation-und-statistik/rechtsvorschriften-lehrplaene/uebersicht-lehrerpruefungen.html

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589–617.

Trempler, K. (2013). KOMPARE – Competent argumentation with evidences: Measurement and modeling in educational sciences and transfer from medical studies In S. Blömeke & O. Zlatkin-Troitschanskaja (Eds.), *The German funding initiative "modeling and measuring competencies in higher education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students* (pp. 78–81). Berlin & Mainz, Germany: Humboldt Universität & Johannes Gutenberg Universität.

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–201. doi:10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203–220. doi:10.1111/j.1745-3984.2000.tb01083.x

Walter, O. (2005). *Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten [Competency assessment in PISA-studies. Simulations for estimating distribution parameters and reliablity]*. Lengerich, Germany: Pabst.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*(1), 92–107.

Watson, J. M., & Callingham, R. A. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3–46.

Wecker, C., Hetmanek, A., & Fischer, F. (2014). *The interplay of domain-specific and domain-general factors in scientific reasoning and argumentation*. Paper presented at the International Conference of the Learning Sciences: Learning and Becoming in Practice, Boulder, CO.

Wei, H. (2008). *Multidimensionality in the NAEP science assessment: Substantive perspectives, psychometric models, and task design*. (Dissertation, University of Maryland). Retrieved from http://drum.lib.umd.edu/bitstream/handle/1903/8048/umi-umd-5194.pdf.

Wenglein, S., Baur, J., Heininger, S., & Prenzel, M. (2015). Kompetenz angehender Lehrkräfte zum Argumentieren mit Evidenz: Erhöht ein Training von Heuristiken die Argumentationsqualität. [Pre-service teachers' evidence-based argumentation competence: Can a training of heuristics improve argumentative quality?]. *Unterrichtswissenschaft, 43*(3), 209–224.

Willison, J., & O'Regan, K. (2007). Commonly known, commonly not known, totally unknown: A framework for students becoming researchers. *Higher Education Research & Development, 26*(4), 393–409.

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

Wilson, M., & Scalise, K. (2006). Assessment to improve learning in higher education: The BEAR Assessment System. *Higher Education, 52*(4), 635–663.

Wirtz, M. A. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann [About the problem of missing data: How the impact of missing information on analysis results can be discovered and reduced]. *Die Rehabilitation, 43*(2), 109–115.

Wirtz, M. A., & Strohmer, J. (2016). Anwendung und Integration qualitativer und quantitativer Forschungsmethoden in der rehabilitationswissenschaftlichen Interventionsforschung. [Application and integration of qualitative and quantitative research methods in intervention studies in rehabilitation research]. *Die Rehabilitation, 55*(03), 191–199. doi:10.1055/s-0042-105940

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Zeuch, N., Förster, N. & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice, 32*(1), 61–70.

Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Toepper, M., & Lautenbach, C. (2016). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen: Ein Überblick zum nationalen und internationalen Forschungsstand* [Measuring academic competencies of university students and absolvents. Overview of the national and international state of research]. Wiesbaden, Germany: Springer.