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-

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

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

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 Studies1)

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 Opho 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-
Assessment of Educational Research Literacy in Higher Education

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, Bensbacher, 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-Paulina (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 parsimonious 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 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 were recruited, and 753 students from three universities 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 \((F_{.05; 1} = 7.447; \eta^2_p = .004)\). 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

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


\(n\) = 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.

<table>
<thead>
<tr>
<th>Heterogeneity AND Elementary School</th>
<th>Heterogeneity OR Elementary School</th>
<th>Heterogeneity NOT Elementary School</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Is it possible to compensate for heterogeneity of elementary school children?</td>
<td>☒</td>
<td>☒</td>
</tr>
<tr>
<td>b) Is it possible to compensate for migration-related disparities in learning conditions of high school students?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>c) Is it possible to compensate for heterogeneity in learning conditions in secondary education?</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

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. quantitative 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:

<table>
<thead>
<tr>
<th></th>
<th>rather applies to A</th>
<th>applies to both A and B</th>
<th>rather applies to B</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) The results give information about the problem in an individual case.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) A general correlation between the attributes &quot;family conflicts&quot; and &quot;school problems&quot; can be deduced.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) The results can be generalized to other adolescents.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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).
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.

<table>
<thead>
<tr>
<th>Competence facets</th>
<th>Winter semester 2012/2013 and summer semester 2013 (study 1)</th>
<th>Summer semester 2014 (study 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unadjusted ($n_i = 226$) adjusted ($n_i = 193$)</td>
<td>unadjusted ($n_i = 29$) adjusted ($n_i = 22$)</td>
</tr>
<tr>
<td>IL</td>
<td>32 (14.2%)</td>
<td>7 (24.1%)</td>
</tr>
<tr>
<td>SL</td>
<td>85 (37.6%)</td>
<td>8 (27.6%)</td>
</tr>
<tr>
<td>ER</td>
<td>109 (48.2%)</td>
<td>14 (48.3%)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Sample</th>
<th>Model</th>
<th>Factors</th>
<th>Final Deviance</th>
<th>n&lt;sub&gt;p&lt;/sub&gt;</th>
<th>AIC</th>
<th>BIC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1, condition a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unadjusted a</td>
<td>1</td>
<td>1 (G)</td>
<td>47590.2</td>
<td>227</td>
<td>48044</td>
<td>49228</td>
<td>49455</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 226)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>47592.0</td>
<td>232</td>
<td>48056</td>
<td>49266</td>
<td>49498</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>47561.2</td>
<td>230</td>
<td>48021</td>
<td>49221</td>
<td>49451</td>
</tr>
<tr>
<td>adjusted b</td>
<td>1</td>
<td>1 (G)</td>
<td>43049.0</td>
<td>194</td>
<td>43437</td>
<td>44449</td>
<td>44643</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 193)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>43052.4</td>
<td>199</td>
<td>43450</td>
<td>44488</td>
<td>44687</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>43020.1</td>
<td>197</td>
<td>43414</td>
<td>44442</td>
<td>44639</td>
</tr>
<tr>
<td>Study 1, condition b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unadjusted a</td>
<td>1</td>
<td>1 (G)</td>
<td>56564.1</td>
<td>227</td>
<td>57018</td>
<td>58202</td>
<td>58429</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 226)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>56425.4</td>
<td>232</td>
<td>56889</td>
<td>58099</td>
<td>58331</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>56402.0</td>
<td>230</td>
<td>56862</td>
<td>58062</td>
<td>58292</td>
</tr>
<tr>
<td>adjusted b</td>
<td>1</td>
<td>1 (G)</td>
<td>51441.6</td>
<td>194</td>
<td>51830</td>
<td>52841</td>
<td>53035</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 193)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>51304.5</td>
<td>199</td>
<td>51703</td>
<td>52740</td>
<td>52939</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>51280.8</td>
<td>197</td>
<td>51675</td>
<td>52702</td>
<td>52899</td>
</tr>
<tr>
<td>Study 2, condition a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unadjusted a</td>
<td>1</td>
<td>1 (G)</td>
<td>18439.2</td>
<td>28</td>
<td>18495</td>
<td>18625</td>
<td>18653</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 27)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>18383.8</td>
<td>33</td>
<td>18450</td>
<td>18602</td>
<td>18635</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>18367.6</td>
<td>31</td>
<td>18430</td>
<td>18573</td>
<td>18604</td>
</tr>
<tr>
<td>adjusted b</td>
<td>1</td>
<td>1 (G)</td>
<td>15871.4</td>
<td>23</td>
<td>15917</td>
<td>16024</td>
<td>16047</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 22)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>15830.4</td>
<td>28</td>
<td>15886</td>
<td>16016</td>
<td>16044</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>15816.6</td>
<td>26</td>
<td>15869</td>
<td>15989</td>
<td>16015</td>
</tr>
<tr>
<td>Study 2, condition b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unadjusted a</td>
<td>1</td>
<td>1 (G)</td>
<td>19917.8</td>
<td>28</td>
<td>19974</td>
<td>20103</td>
<td>20131</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 27)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>19807.7</td>
<td>33</td>
<td>19874</td>
<td>20026</td>
<td>20059</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>19796.0</td>
<td>31</td>
<td>19858</td>
<td>20001</td>
<td>20032</td>
</tr>
<tr>
<td>adjusted b</td>
<td>1</td>
<td>1 (G)</td>
<td>17171.6</td>
<td>23</td>
<td>17218</td>
<td>17324</td>
<td>17347</td>
</tr>
<tr>
<td>(n&lt;sub&gt;i&lt;/sub&gt; = 22)</td>
<td>2</td>
<td>3 (IL, SL, ER)</td>
<td>17097.0</td>
<td>28</td>
<td>17153</td>
<td>17283</td>
<td>17310</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4 (G, IL, SL, ER)</td>
<td>17087.4</td>
<td>26</td>
<td>17139</td>
<td>17260</td>
<td>17286</td>
</tr>
</tbody>
</table>

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.

ni = number of test items included; n<sub>p</sub> = 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

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

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

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

Note. Study 1: winter semester 2012/2013 and summer semester 2013, N = 1360, n = 193. Study 2: summer semester 2014, N = 753, n = 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 Opho ff, 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 Opho ff, & 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. Olehmovica 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 & Mauder, 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 helpful support concerning questions about the R-package Test Analysis Modules (TAM).
References


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


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


Schladitz, S., Rott, B., Winter, A., Wischgoll, A., Groß Ophoff, J., Hosenfeld, I., Wittwer, J. (2013). LeSeEd – 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-
Assessment of Educational Research Literacy in Higher Education


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. 71–74). Berlin & Mainz, Germany: Humboldt Universität & Johannes Gutenberg Universität.


