Studia paedagogica
vol. 26, n. 4, 2021
www.studiapaedagogica.cz
https://doi.org/10.5817/SP2021-4-2

ASSESSMENT OF GERMAN AND AUSTRIAN STUDENTS' EDUCATIONAL RESEARCH LITERACY: VALIDATION OF A COMPETENCY TEST BASED ON CROSS-NATIONAL COMPARISONS

JANA GROß OPHOFF,
CHRISTINA EGGER

Abstract

Educational Research Literacy (ERL) is the ability to access, comprehend, and consider scientific information and to apply the resulting conclusions to problems connected with educational decisions. It is crucial for the process of data-based decision making and–corresponding to the consecutive phases–defined as the conglomeration of different facets of competence, including information literacy, statistical literacy, and evidence-based reasoning. However, the engagement with research in educational contexts appears to have some difficulties. This is even more remarkable as the state of knowledge about actual teacher competency levels remains unsatisfactory, even though test instruments for assessing research literacy have been developed in recent years. This paper addresses the question of whether such a test developed in the specific context of German study programs in (teacher) education can be applied to other national contexts, in this case to Austrian teacher education. An investigation of the construct validity under consideration of the psychometric structure and group differences on item level is necessary for ensuring the fairness of cross-national comparisons. Based on multidimensional item response theory models, samples from Germany (n = 1360 students, 6 universities) and Austria (n = 295 students, 2 universities) are investigated in terms of measurement invariance between the two countries. A comparable psychometric structure and at least partial measurement invariance with no particular advantage for either sample could be demonstrated. This is an indication that the presented test instrument can be validly applied to assess the research literacy of teacher training students in both countries.

Keywords

Educational Research Literacy, teacher education, construct validity, measurement equivalence, differential item functioning

Introduction

As early as 1999, Davies stated that educational professionals at all levels should be able to (a) to pose answerable questions; (b) search for relevant information; (c) read and critically appraise evidence; (d) evaluate; and (e) use the resulting conclusions for educational decision making. These requirements correspond to the stages of research engagement in the sense of a complex, cognitive, knowledge-based problem-solving cycle. Thus, it is not surprising that corresponding process descriptions can be found in conceptual frameworks of data-based decision making (e.g., Groß Ophoff & Cramer, in press; Mandinach et al., 2008; Marsh, 2012; Schildkamp & Kuiper, 2010; Schildkamp & Poortman, 2015; Schratz et al., 2018). These models consider teachers' competent engagement with and the use of research in the various forms of data and evidence available to teachers (cf., Wiesner & Schreiner, 2019) as crucial for quality improvement and professionalization in educational practice. Accordingly, there is some evidence that if educators engage with evidence to make or change decisions, embark on new courses of action, or develop new practices, this can have a positive impact on both teaching and learning (Bach et al., 2014; Cain, 2015; Richter et al., 2014; van Geel et al., 2016). However, there is evidence that teachers still struggle to transform data from performance tests, and also from classroom records, classroom assessments, program descriptions, and school statistics, etc. into useful knowledge (Groß Ophoff & Cramer, in press; Hamilton & Reeves, 2021; Schildkamp & Lai, 2013). Instead, teachers appear to rely on intuition, which is prone to bias and mistakes (Dunn et al., 2019; Fullan, 2005). Even attempts to develop the capacity of school leaders and practitioners to engage in reflective problem solving, such as Research Learning Networks or Data Teams (Brown et al., 2017; Mintrop & Zumpe, 2019), seem to fail in their attempts to facilitate deep research engagement. Against this backdrop, this paper addresses the issue of the assessment of the necessary and crucial competencies that should enable teachers and educational practitioners in general to engage and use research deliberately and (more) systematically.

Theoretical Background

In the field of educational assessment, the widely called-for research-related competencies include Educational Research Literacy (ERL, Groß Ophoff, Schladitz, et al., 2017; cf., Shank & Brown, 2007). This is conceptually related to assessment literacy (referring to the selection and use of student assessments, cf., DeLuca et al., 2016), data literacy (referring to drawing instructional conclusions from statistical information, cf., Mandinach, 2012; van Geel et al., 2017a), and statistical literacy (SL, referring to organizing/working with different data representations and understanding statistical concepts, cf., Ben-Zvi & Garfield, 2004; Watson & Callingham, 2003). In order to capture the research cycle as a whole, the concept of ERL also incorporates concepts from adjacent research fields, including information literacy (IL, referring to formulating research questions and information searches, e.g., Blixrud, 2003) and evidence-based reasoning (ER, referring to interpreting and critically evaluating evidence, e.g., Kuhn et al., 2008; Halonen, 2008).

However, despite the global movement toward accountability, evaluation, and assessment in education (DeLuca & Johnson, 2017) and the (theoretically assumed) importance of educational practitioners' proficiency in the engagement with research (see above), the state of knowledge about actual competency levels is unsatisfactory, and not only in Germany and Austria. However, there is some evidence that German in-service teachers are less proficient in ERL than pre-service teachers, even though the required abilities can be imparted or fostered during initial training or through professional development (e.g., Kittel et al., 2017). Accordingly, university (teacher) education is viewed as central because it allows connecting research and teaching (Healey, 2005). Because of the reorganization and change processes associated with the Bologna Reform (e.g., German Federal Ministry of Education and Research, 2015), a theoretical and empirical foundation for developing and implementing sustainable and psychometrically sound measures for quality assurance and development is regarded as crucial (Blömeke & Zlatkin-Troitschanskaja, 2013). In particular, psychometrically sound test instruments are expected to support the criterion-referenced interpretation with regard to the aspired competencies–which in turn can stimulate curriculum development and facilitate feedback about learning goals and gains (Wilson & Scalise, 2006) in initial education and as part of continuing teacher education.

Research literacy is usually assessed based on self-reports (e.g., Braun et al., 2008; Ntuli & Kyei-Blankson, 2016), but in general, correlations between subjective and objective competency measures are rather low (Lowman & Williams, 1987). Empirical approaches via assigned test instruments can be found, but have been scarce and psychometrically weak (e.g., Reeves & Honig, 2015). Recently developed test instruments focus either on particular steps of the research cycle (e.g., ER: Münchow et al., 2019; SL: Zeuch et al., 2017), or were developed in the context of specific interventions (e.g., Ebbeler et al., 2017; van Geel et al., 2016). Regarding the investigation of the psychometric structure, more often than not, one-dimensional models are applied without further comparison to other theoretically plausible multidimensional models (e.g., Watson & Callingham, 2003; van Geel et al., 2016). As one approach to investigating construct validity (Cronbach & Meehl, 1955), Groß Ophoff, Schladitz, et al. (2017) compared theoretically plausible one- and multi-dimensional models of a test instrument for the (more comprehensive) assessment of ERL based on a sample of 1360 students at six German universities. Even though it could be demonstrated that ERL consists of one generic factor of ERL and three secondary factors representing specific aspects in relation to the requirements of the research cycle (IL, SL, ER), the authors recommended applying a one-dimensional model because–due to the dominance of the general factor–essential unidimensionality can be assumed (Stout, 1987). Additionally, there is some evidence that even though social sciences share a certain methodological repertoire (Dietrich et al., 2015), the different research traditions represented in study programs involved in teacher education (e.g., sociology, educational science, psychology) appear to have a differential impact on performance in comprehensive assessments of research competency (Gess et al., 2017). In line with conceptual frameworks of data-based decision making (see above), the acquisition of competencies is shaped by the research-related opportunities-to-learn during initial and continuing teacher education (based on the institution- and discipline-specific curriculum) and also by its national and cultural contexts (Larcher & Oelkers, 2004). This perspective has been adopted in the current contribution.

For example, the educational systems in Germany and Austria (and Switzerland, for that matter) share cultural and linguistic commonalities (Gonon, 2011). In recent history, both countries were faced with an empirical shift in their education systems after disappointing results in international large-scale assessments became public in the early 2000s (Altrichter et al., 2005; Bos et al., 2010). In the aftermath, as early as 2004, research literacy was explicitly identified as a requirement in the so-called Standards for Teacher Education by the German Standing Conference of the Ministers of Education and Cultural Affairs. Accordingly, teachers in training should be able to consider and evaluate evidence from educational research and practicing teachers should be able to use evidence-based insights for instructional and school development. In Austria, reforms came to fruition later, especially as stakeholders in education policy did not present a united front (Olano, 2010). As late as 2010, the Austrian Federal Ministry of Science and Research and the Federal Ministry for Education, Arts and Culture (2010) published an expert view on the future of pedagogical professions in which the recommendation was expressed that science and research need to be established as constitutive elements of teacher education. Further reforms in teacher education followed in later years, like the legal adoption of a reform in 2013 (Hofmann et al., 2020) asserting that all teachers in training must obtain an academic degree (bachelor's or master's degree), and renouncing the previously parallel organization of teacher education in universities of education (UE, German: Pädagogische Hochschulen; with a focus on primary and lower secondary education) and universities (with a focus on higher secondary education) in favor of founding development networks or clusters in which universities and UE collaborate.

Research Questions

Against this backdrop, this paper addresses the question of whether a test instrument developed in the specific context of German study programs in (teacher) education can be applied to other national contexts, in this case to Austrian teacher education. This approach to investigating construct validity under consideration of the psychometric structure and group differences on item level (Cronbach & Meehl, 1955) is a necessary step in ensuring the fairness of cross-national comparisons (Davidov et al., 2014; Förster et al., 2015).

For this purpose, results about the dimensionality of ERL in the large-scale German study (e.g., Groß Ophoff, Wolf, et al., 2017) are compared to a study at two Austrian UE (Haberfellner, 2016). In both studies, the same ERL test instrument was used. According to Prenzel et al. (2007), probabilistic test theory, which is the basis for the reported analyses in this paper, makes it possible to validate theoretically plausible assumptions about the dimensional structure of a construct (e.g., by comparing competing models). Accordingly, for the data from the German sample (Study 1), a bifactor model (Model 3, see 4.3) with one dominant general factor and the three secondary factors (IL, SL, ER) turned out to be the best fit (Groß Ophoff, Wolf, et al., 2017). This model served as acceptable compromise between the one-dimensional model (Model 1) and the three-dimensional model (Model 2 with the subdimensions IL, SL, ER) that were applied in preliminary analyses (e.g., Haberfellner, 2016; Schladitz et al., 2015). These findings serve as a reference for the analysis of the Austrian sample in this paper. As the invariance of the measurement instrument is crucial for the valid comparison of samples from different countries (e.g., Davidov et al., 2014), the following question will be pursued:

1. Can the psychometric structure of the ERL test instrument for the sample of German students in Study 1 also be applied to the sample of Austrian students in Study 2 (Model 3), and can configural invariance therefore be assumed? If not, which of the two other theoretically plausible models (Model 1, Model 2) fits better?
2. Are different probabilities for a correct response in single items identifiable (so-called differential item functioning; DIF)? If so, is one of the samples consistently disadvantaged (uniform DIF) or does this vary across the item sample (non-uniform DIF)?

Methods

Analyses were conducted utilizing data sets from two studies: the first from the large-scale main study in Germany (Study 1: winter semester 2012/2013 and summer semester 2013), and the second from a study at two Austrian UE (German: Pädagogische Hochschulen = UE) in the summer semester of 2015 (Study 2). In both studies, participants were recruited upon request in lectures (convenience samples). Participation was voluntary and anonymous.

Data collection and samples

In Study 1 (see Table 1), 1360 students in the field of educational science at six German higher education institutions from five federal states were investigated between 2012 and 2013. Because of the German federal constitution, the federal states are predominantly responsible for education, science, and culture, but cross-nationally coordinate and collaborate in education and training (to some extent) through the Standing Conference established in 1948 (Standing Conference, 2019, 2020). The sample includes one UE in Baden-Württemberg, and one university (reformed former UE) in Rhineland-Palatinate that both are rather small universities with a strong focus on educational science and related disciplines as well as on subject-related didactics. In other federal states, teacher education institutions were integrated into the educational science departments of state universities by the 1970s (Meissner et al., 2012). This is the case for the other four large universities in this sample that offer a wide range of study programs and are characterized by a strong research orientation. In teacher training, these comprehensive universities typically tend to focus on subject-related studies. In this study, teacher training students (for all school forms) represented the largest group, followed by educational studies students (23%), and other study programs (e.g., early education, health education, educational psychology).

Table 1
Descriptive statistics of the samples from Study 1 (five German states, winter semester 2012/2013 and summer semester 2013) and Study 2 (two development networks: summer semester 2015)

	Germany (*Study 1*): winter semester 2012/2013 and summer semester 2013	Austria (*Study 2*): summer semester 2015
N	1360 students	295
Age, *M* (*SD*)	22.9 years (3.95)	22.9 years (4.32)
Gender (% female)	75.9%	77.6%
Teacher Training students	62%	100%

Note. Abbreviations: *N* = number of study participants.

Study 2 investigated 295 teacher training students (primary and lower secondary education) from two Austrian UE from the cluster "West" (Tirol, Vorarlberg) and "Mitte" (Salzburg, Upper Austria). At the time of the study, teacher education for primary and lower secondary schools was located at UE that were established as late as 2007 from post-secondary schools (German: Pädagogische Akademien). To this day, Austrian UE are not authorized to award doctoral and postdoctoral degrees, and the link between research and teaching is by no means a given for teaching staff (Haberfellner, 2016; Hofmann et al., 2020).

Test instrument and booklet design

The main focus of the research program in Study 1 was on the development of a test instrument for the assessment of ERL (see Table 2) covering the steps of the research process, such as search strategies for problem-specific research information, the comprehension of different types of academic documents, the formulation of adequate research questions (IL), the analysis and interpretation of descriptive statistics (SL), and the critical evaluation of research-based assumptions (ER; cf., Groß Ophoff, Wolf, et al., 2017). The resulting item pool was reviewed by content experts, pre-tested comprehensively, and subsequently deployed with the goal of test standardization between 2012 and 2013 (see Table 1). During implementation, 40 minutes were allotted by the test administrators to complete the ERL test. In the remaining 20 minutes, participants were asked to provide personal and professional background information, and further characteristics were surveyed. During data analysis in Study 1, poor fitting items (0.80 ≥ Infit/Outfit ≥ 1.20, cf., Adams & Wu, 2002) and items with low discrimination (r < 0.20) were excluded. The foundation of the results reported here is the reduced item pool of 193 items (119 stems, see Table 2).

Table 2

Distribution of test items to the competence facets information literacy, statistical literacy, and evidence-based reasoning in the standardization study (Germany) and the study in summer semester 2015 (Austria)

	Germany (DE: Study 1): winter semester 2012/2013 and summer semester 2013	Austria (AT: Study 2): summer semester 2015
Competence facets		
IL	30 (15.5%)	8 (20.0%)
SL	71 (36.8%)	14 (35.0%)
ER	92 (47.4%)	18 (45.0%)

Note. Abbreviations: IL = information literacy; SL = statistical literacy; ER = evidence-based reasoning; n_i = number of test items.

In contrast to Study 1, the focus of Study 2 was on investigating the effect of the subjective value of research on pre-service teachers' research-oriented stance and on their level of ERL (Haberfellner, 2016). The 40 test items (referring to 17 stems, see Table 2) were selected from the item pool from Study 1 and then arranged in a single test booklet set up for a processing time of 40 minutes. Again, personal and professional background information were collected, and research-related attitudes were assessed. The test booklet for Study 2 had to be compiled before the data analysis in Study 1 was concluded. Therefore, no standardized parameter estimates were available for six of the selected items because they were excluded from analysis in Study 1 (see above). In Study 2, all items showed good item fit and were retained in the separate investigation of the dimensional structure of ERL reported here. These items could not be used to investigate differential item functioning (DIF, see 4.3), testing for (partial) measurement equivalence. The same applies to eight other items that were slightly modified for Study 2. Therefore, the in-depth analysis of the item-by-country interaction was based on 26 items.

Statistical analysis

Psychometric models popular in the field of competency assessment are based on item response theory (IRT), which rests upon stringent statistical assumptions (i.e., monotonicity, local independence, and unidimensionality). Multidimensional IRT models (Hartig & Höhler, 2009) assume that several latent dimensions are represented by item clusters. But it has been questioned whether the assumption of strict unidimensionality is applicable to, for example, educational and psychological assessment where, in addition to one dominant latent trait, other minor latent factors likely influence participant responses (e.g., Gustafson, 2001). Bifactor models are a solution to this, as they allow each item response to be explained by both a dominant factor in the sense of a common latent trait (e.g., ERL), and additional, orthogonal (therefore uncorrelated) factors caused by "parcels" of items drawing from similar aspects of the underlying traits (Reise et al., 2010).

As mentioned above, valid comparisons between groups–like the samples from Study 1 and Study 2–require cross-national invariance of the measurement instrument (Tay et al., 2015). The identification of a comparable dimensional structure for the ERL instrument ("configural invariance") was the first step in warranting comparability. Therefore, in each of the two samples (Study 1, Study 2), three competing models (see 3) were compared with the R package Test Analysis Modules (TAM, Kiefer et al., 2016). The best fitting model was identified separately for each sample based on the lowest values in the information criteria AIC, BIC, and CAIC (Schermelleh-Engel et al., 2003). The precision of person estimates 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 (Bond & Fox, 2007). This coefficient is comparable with Cronbach's α, for which values of at least 0.55 are deemed satisfactory for group comparisons (Rost, 2013). For multidimensional constructs like the bifactor model, Green and Yang (2009) recommended reporting Omega (ω) as a model-based reliability estimate that combines higher-order and lower-order factors, and Omega-hierarchical (ωh) as model-based reliability estimate of one target construct with others removed.

To gain further insights into measurement equivalence (or the lack thereof) of single items, DIF was investigated. To this end, group specific item parameters were compared based on the deviation of the group mean from the overall mean in relation to the standard error (Critical Ratio, cf., Holland & Wainer, 1993). Accordingly, values for a certain item below $z = -1.96$ or above $z = 1.96$ indicate meaningful DIF (Wu et al., 2007). In this case, respondents with the same proficiency level, but from different countries, showed different probabilities for a correct response in an item (Wirtz & Böcker, 2017). However, emerging DIF should be interpreted with caution here because smaller samples lead to higher standard errors, thus more frequently to significant results. This is particularly the case in Study 1, where single items were usually assigned to approximately 200 students due to the applied incomplete block design (Groß Ophoff, Wolf, et al., 2017).

Results

At first glance, the samples from Study 1 and Study 2 appear to demonstrate a different dimensionality of ERL (see Table 3). In Study 1, the bifactor model solution in Model 3 shows better fit than the one- or the three-dimensional models because the corresponding values of AIC, BIC, and CAIC were lowest. The information criteria values of the one-dimensional and the bifactor model were closer to each other than to the three-dimensional model. This is the same in Study 2, even though only the AIC indicates the four-dimensional model as better-fitting, whereas the BIC- and CAIC-values favored the more parsimonious one-dimensional model of ERL. Overall, the model results from both samples indicate that the three secondary factors of IL, SL, and ER can be distinguished from a general factor of ERL. Although more pronounced in the Study 1 sample, the general factor in Model 3 was dominant in both samples (DE: ωh = 0.85; AT: ωh = 0.65). Accordingly, it is reasonable to apply a one-dimensional model without further differentiation of the three competence facets (Groß Ophoff, Wolf, et al., 2017b). For the one-dimensional model, the reliability of the test instrument was found to be satisfactory for both the German (EAP-reliability = 0.61, cf., Böttcher-Oschmann et al., 2019) and the Austrian sample (EAP-reliability = 0.59).

Table 3
Goodness-of-fit statistics for competing models in Study 1 and Study 2

Sample	n_i	Model	Factors	Final Deviance	n_p	AIC	BIC	CAIC
Study 1: DE	193	1	1 (G)	43,049.0	194	43,437	44,449	44,643
		2	3 (IL, SL, ER)	43,052.4	199	43,450	44,488	44,687
		3	4 (G, IL, SL, ER)	43,020.1	197	**43,414**	**44,442**	**44,639**
Study 2: AT	40	1	1 (G)	10,760.7	41	10,843	**10,994**	**11,035**
		2	3 (IL, SL, ER)	10,742.7	46	10,835	11,004	11,050
		3	4 (G, IL, SL, ER)	10,744.5	44	**10,832**	10,995	11,039

Note. Study 1 (Germany): winter semester 2012/2013 & summer semester 2013. *Study 2* (Austria): summer semester 2015. Sample size: *N (Study 1)* = 1360; *N (Study 2)* = 295. abbreviations: 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; ER = evidence-based reasoning; The parameters of the respective best fitting solution are indicated in bold.

On closer inspection of the 26 test items included in the DIF analysis (see 4.3), 12 items showed no meaningful DIF between the two samples. In Table 4, the results for the remaining 14 items are reported. Critical Ratio values (last column) below z = −1.96 indicate that the teacher training students in Study 1 showed a higher probability for a correct response than those in Study 2 (upper half of Table 4), which is the case for six items; conversely, values above z = 1.96 indicate an advantage for participants in Study 2 (eight items, see lower half of Table 4). In the third and fourth column from left, task content and the required competencies are briefly stated. It should be stressed that for item 6.1, the slight advantage for the Austrian sample might be explained by the compilation of the test booklet. In Study 2, this item was preceded by item 5, referring to the same graph; in Study 1, these two items were located in different test booklets. The obvious assumption is that the close reading required for the solution of item 5 lead to a slight advantage in this item. But overall, a mixed picture emerges.

Table 4

Overview of items with Differential Item Functioning in Study 1 and Study 2

In favor of …	Item position (Study 2)	Task content	Required competencies	M (Study 1)	SE (Study 1)	Critical Ratio
Study 1	10	Study abstract	Identification of adequate follow-up research question	−0.643	0.082	−7.84
	11	Literature search	Identification of suitable search terms	−0.246	0.076	−3.24
	16.4	Comparison of two study abstracts	Evaluation of study designs	−0.312	0.079	−3.95
	16.5		Identification of study with control group	−0.331	0.079	−4.19
	21.3	Bar chart (degree aspiration of male vs. female students)	Recognition of inadmissible conclusion	−0.513	0.097	−5.29
	21.2		Calculation of percentages for appraising a statement	−0.273	0.081	−3.37
Study 2	4.1	Description of different research procedures	Assessment of suitability for research objective	0.592	0.081	7.31
	4.2			0.408	0.084	4.86
	4.3			0.201	0.084	2.39
	6.1	Integrated bar chart	Graph interpretation	0.163	0.075	2.17
	8	Venn diagram	Interpretation of intersections	0.352	0.099	3.56
	16.2	see 16.4 (above)	Appraisal of conclusions	0.191	0.078	2.45
	16.6			0.228	0.082	2.78
	19	bibliographical reference	Identification of source	0.423	0.078	5.42

Note. Study 1 (Germany): winter semester 2012/2013 & summer semester 2013. *Study 2* (Austria): summer semester 2015. Abbreviations: IL = information literacy; SL = statistical literacy; ER = evidence-based reasoning. All items reported show significant DIF (Critical Ratio below z = −1.96 or above z = 1.96).

Conclusions

Even though these reported results appear somewhat inconclusive with a view to the dimensionality of ERL, they perpetuate the previously described structural ambiguity of the test instrument (Groß Ophoff, Wolf, et al., 2017). Because of the dominance of the general factor in Model 3 in both samples (DE: $\omega_h = 0.85$; AT: $\omega_h = 0.65$), the recommendation to use a one-dimensional model of ERL (Model 1) for the assessment and feedback of proficiency on the individual level (research question 1, see 3) could be substantiated.

Overall, the results indicate that the presented ERL test can be validly applied to assess the research literacy of teacher training students in both countries, even though it is worthwhile to take Differential Item Functioning into account. The DIF analysis of the two samples further revealed that at least partial equivalence can be assumed (research question 2), even though the issue of whether the identified violations are problematic for meaningful comparisons is still controversial (Davidov et al., 2014). Nevertheless, the identification of non-uniform DIF indicates that neither of the two samples was consistently disadvantaged. Given that both studies are based on convenience samples (which is a common challenge for research in higher education, cf., Zlatkin-Troitschanskaia et al., 2016), items with DIF might–interpreted with due caution–hint at some advantage in research-methodological issues for the sample in Study 1, and in appraising research-based conclusions for Study 2. But it should be remembered that even though more items showed an unexpected higher probability for a correct response for the students in Study 2, they showed an overall lower proficiency in ERL ($b_{\text{Study2}} = -.22$; 95%-CI: −.30, −.14) than teacher training students in Study 1 (NLA = 841). However, particularly in Study 1, the ratio of persons per item was comparatively small due to the incomplete block design. To gain a better understanding of the reasons for the identified DIF (benign vs. adverse DIF, cf., Gierl, 2005), larger, specifically selected item samples need to be investigated based on larger samples, and curricular content experts should be involved.

It should be noted, too, that it was not necessary to translate the test items in the two studies here. The transfer to other cultural contexts and the necessary translation to ensure linguistic and cultural equivalence will probably present greater challenges (e.g., Grisay et al., 2007). For example, there are currently translations for a selection of ERL test items either available (English, Arabic) or in the making (Italian, Spanish). But due to the use of the translated ERL tests for course or curricular evaluations in specific higher education institutions and the resulting small samples, the translated versions have not yet been analyzed with regard to measurement equivalence.

In-depth analysis on a meso-level revealed differences between the institutions included in the two studies here. For example, the more proficient teacher training students in Study 1 were located at large German universities with a traditionally strong research orientation, whereas students with the lowest proficiency came from a university that did not explicitly identify ERL as a study objective in the curriculum at that time (Groß Ophoff, Schladitz, et al., 2017). In Study 2, both institutions offered only introductory research-based courses (scientific working methods, applied research, and evaluation), which is probably why no significant differences in ERL between the two emerged. Presumably, these differences are related to the embedding and amount of research in teacher education study programs. While knowledge about the extent of research-orientation in Austrian education is still rather limited (Jesacher-Roessler & Kemethofer, in press), more is known about the current research-related practices in German teacher education (Groß Ophoff & Cramer, in press): Evidently mainly research-led (focused on engagement with research data) or—to a lesser extent—research-oriented courses (focused on imparting research methods, e.g., Rueß et al., 2016; Stelter & Miethe, 2019) are available. Inquiry-based courses in which students are scaffolded to absolve certain phases or full research projects have been established in recent years, particularly as part of long-term school internships (e.g., Ulrich & Gröschner, 2020). But findings about the intended (research-related) effects have thus far been rather sobering (e.g., van Ophuysen et al., 2017). Based on the reconstructive analysis of inquiry-based course concepts in teacher education, Katenbrink and Goldmann (2020) pointed out that rather superficial "one fits all"-concepts appear to dominate in German initial teacher education, in which practical procedures are trained and inquiry is loosely imparted as the evaluation of educational practices.

Further limitations of the study presented here are that the ERL test strongly (but not exclusively) focuses on quantitative-methodological topics. Furthermore, the presented ERL test operationalizes only a subsidiary, that is cognitive, aspect of research competence. In recent years, there has been an increased awareness that affective-motivational factors also play an important role for the depth of engagement with research information (Wessels et al., 2018), which highlights the importance of further research on meso- (courses in teacher training) and micro-level (competency development of pre- and in-service teachers). This might shed light on the much-needed advanced understanding of how to support or facilitate competent research engagement in teacher education (cf., Brown et al., 2021). According to Katenbrink and Goldmann (2020), inquiry-based learning in particular, with a focus on the assumption of the fundamental and unresolvable difference between theory and practice (concepts of difference), has the potential to empower teachers to reflect on their own educational practice with professional distance (Cramer et al., 2019; Helsper, 2016) and to use research information as an opportunity for deep learning or even conceptual change (Gregoire, 2003), and also to convince them about the usefulness of research for quality development in education (Prenger & Schildkamp, 2018).

References

Adams, R. J., & Wu, M. (2002). *PISA 2000 Technical Report.* Organisation for Economic Cooperation and Development. https://www.oecd.org/pisa/data/33688233.pdf

Altrichter, H., Brüsemeister, T., & Heinrich, M. (2005). Merkmale und Fragen einer Governance-Reform am Beispiel des österreichischen Schulwesens. *Österreichische Zeitschrift für Soziologie*, *30*(4), 6–28. https://doi.org/10.1007/s11614-006-0063-0

Bach, A., Wurster, S., Thillmann, K., Pant, H. A., & Thiel, F. (2014). Vergleichsarbeiten und schulische Personalentwicklung—Ausmaß und Voraussetzungen der Datennutzung. *Zeitschrift für Erziehungswissenschaft*, *17*(1), 61–84. https://doi.org/10.1007/s11618-014-0486-5

Ben-Zvi, D., & Garfield, J. (2004). The challenge of developing statistical literacy, reasoning and thinking (1st ed.). Springer. https://doi.org/10.1007/1-4020-2278-6

Blixrud, J. C. (2003). Project SAILS: Standardized assessment of information literacy skills. In G. J. Barret (Ed.), *ARL: A bimonthly report on research library issues and actions from ARL, CNI, and SPARC. Special double issue on new measures. Number 230-231, October-December 2003* (pp. 18–19). Association of Research Libraries. https://files.eric.ed.gov/fulltext/ED498293.pdf

Blömeke, S., & Zlatkin-Troitschanskaia, O. (2013). Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs (KoKoHs Working Papers No. 1). Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor. https://www.kompetenzen-im-hochschulsektor.de/files/2018/05/KoKoHs_WP1_Bloemeke_Zlatkin-Troitschanskaia_2013_.pdf

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Erlbaum.

Bos, W., Postlethwaite, T. N., & Gebauer, M. M. (2010). Potenziale, Grenzen und Perspektiven internationaler Schulleistungsforschung. In R. Tippelt & B. Schmidt (Eds.), *Handbuch Bildungsforschung* (3rd ed., pp. 275–295). Springer.

Böttcher-Oschmann, F., Groß Ophoff, J., & Thiel, F. (2019). Validierung eines Fragenbogens zur Erfassung studentischer Forschungskompetenzen über Selbsteinschätzungen – Ein Instrument zur Evaluation forschungsorientierter Lehr-Lernarrangements. *Unterrichtswissenschaft*, *47*(4), 495–521. https://doi.org/10.1007/s42010-019-00053-8

Braun, E., Gusy, B., Leidner, B., & Hannover, B. (2008). Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp). *Diagnostica*, *54*(1), 30–42. https://doi.org/10.1026/0012-1924.54.1.30

Brown, C., Poortman, C., Gray, H., Groß Ophoff, J., & Wharf, M. (2021). Facilitating collaborative reflective inquiry amongst teachers: What do we currently know? *International Journal of Educational Research*, *105*, 101695. https://doi.org/10.1016/j.ijer.2020.101695

Brown, C., Schildkamp, K., & Hubers, M. D. (2017). Combining the best of two worlds: A conceptual proposal for evidence-informed school improvement. *Educational Research*, *59*(2), 154–172. https://doi.org/10.1080/00131881.2017.1304327

Bundesministerium für Unterricht, Kunst und Kultur, & Bundesministerium für Wissenschaft und Forschung [Federal Ministry for Education, the Arts and Culture & Federal Ministry of Science and Research]. (2010). *LehrerInnenbildung NEU – Die Zukunft der pädagogischen Berufe: Die Empfehlungen der ExpertInnengruppe* [Teacher Education NEW – The future of pedagogical professions]. https://www.qsr.or.at/dokumente/1870-20140529-092820-Empfehlungen_der_ExpertInnengruppe_Endbericht_092010_2_Auflage.pdf

Cain, T. (2015). Teachers' engagement with research texts: Beyond instrumental, conceptual or strategic use. *Journal of Education for Teaching*, *41*(5), 478–492. https://doi.org/10.1080/02607476.2015.1105536

Cramer, C., Harant, M., Merk, S., Drahmann, M., & Emmerich, M. (2019). Meta-Reflexivität und Professionalität im Lehrerinnen- und Lehrerberuf. *Zeitschrift für Pädagogik*, *65*(3), 401–423.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, *47*(2), 108–121. https://doi.org/10.1111/1467-8527.00106

DeLuca, C., & Johnson, S. (2017). Developing assessment capable teachers in this age of accountability. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 121–126. https://doi.org/10.1080/0969594X.2017.1297010

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. https://doi.org/10.1007/s11092-015-9233-6

de Olano, D. (2010). Gewinner, Verlieren und Exoten - PISA in sieben weiteren Staaten. In P. Knodel, K. Martens, D. de Olano, & M. Popp (Eds.), *Das PISA-Echo. Internationale Reaktionen auf die Bildungsstudie* (pp. 251–300). Campus.

Dietrich, H., Zhang, Y., Klopp, E., Brünken, R., Krause, U.-M., Spinath, F. M., Stark, R., & Spinath, B. (2015). Scientific competencies in the social sciences. *Psychology Learning & Teaching*, *14*(2), 115–130. https://doi.org/10.1177/1475725715592287

Dunn, K. E., Skutnik, A., Patti, C., & Sohn, B. (2019). Disdain to acceptance: Future teachers' conceptual change related to data-driven decision making. *Action in Teacher Education*, *41*(3), 193–211. https://doi.org/10.1080/01626620.2019.1582116

Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (2017). The effects of a data use intervention on educators' satisfaction and data literacy. *Educational Assessment, Evaluation and Accountability*, *29*(1), 83–105. https://doi.org/10.1007/s11092-016-9251-z

Federal Ministry of Education and Research Germany. (2015). *Bericht der Bundesregierung über die Umsetzung des Bologna-Prozesses 2012–2015 in Deutschland.* https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2015/2015_02_12-NationalerBericht_Umsetzung_BolognaProzess.pdf

Förster, M., Zlatkin-Troitschanskaia, O., Brückner, S., Happ, R., Hambleton, R. K., Walstad, W. B., Asano, T., & Yamaoka, M. (2015). Validating test score interpretations by cross-national comparison: Comparing the results of students from Japan and Germany on an American test of economic knowledge in higher education. *Zeitschrift Für Psychologie*, *223*(1), 14–23. https://doi.org/10.1027/2151-2604/a000195

Fullan, M. (2005). The meaning of educational change: A quarter of a century of learning. In A. Lieberman (Ed.), *The roots of educational change: International handbook of educational change* (1st ed., pp. 202–216). Springer. https://doi.org/10.1007/1-4020-4451-8_12

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), 11–36. https://doi.org/10.25656/01:14895

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, *24*(1), 3–14. https://doi.org/10.1111/j.1745-3992.2005.00002.x

Gonon, P. (2011). Die Bedeutung des internationalen Arguments in der Lehrerbildung. *Beiträge zur Lehrerbildung, 29*(1), 20–26. https://doi.org/10.25656/01:13763

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. https://doi.org/10.1007/s11336-008-9099-3

Gregoire, M. (2003). Is it a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change. *Educational Psychology Review*, *15*(2), 147–179. https://doi.org/10.1023/A:1023477131081

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249–266.

Groß Ophoff, J., & Cramer, C. (in press). The engagement of teachers and school leaders with data, evidence and research in Germany. In C. Brown & J. R. Malin (Eds.), *The Emerald international handbook of evidence-informed practice in education: Learning from international contexts.* Emerald.

Groß Ophoff, J., Schladitz, S., & Wirtz, M. (2017). Differences in research literacy in educational science depending on study program and university. In J. Domenech i de Soria, M. C. Vincent-Vela, E. de la Poza, & D. Blazquez (Eds.), *Proceedings of the HEAd'17. 3rd international conference on higher education advances* (pp. 1193–1202). Congress UPV. http://dx.doi.org/10.4995/HEAD17.2017.6713

Groß Ophoff, J., Wolf, R., Schladitz, S., & Wirtz, M. (2017). 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*(2), 37–68. https://doi.org/10.25656/01:14896

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* (1st ed., pp. 87–111). Routledge.

Haberfellner, C. (2016). Der Nutzen von Forschungskompetenz im Lehramt: Eine Einschätzung aus der Sicht von Studierenden der Pädagogischen Hochschulen in Österreich. Klinkhardt.

Halonen, J. S. (2008). Measure for measure: The challenge of assessing critical thinking. In D. S. Dunn, J. S. Halonen, & R. A. Smith (Eds.), *Teaching critical thinking in psychology: A handbook of best practices* (1st ed., pp. 61–76). Wiley-Blackwell.

Hamilton, V. M., & Reeves, T. D. (2021). Relationships between course taking and teacher self-efficacy and anxiety for data-driven decision making. *The Teacher Educator,* 1–19. https://doi.org/10.1080/08878730.2021.1965682

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

Healey, M. (2005). Linking research and teaching: Exploring disciplinary spaces and the role of inquiry-based learning. In R. Barnett (Ed.), *Reshaping the University: New Relationships between Research, Scholarship and Teaching* (1st ed., pp. 67–78). Open University Press.

Helsper, W. (2016). Lehrerprofessionalität – der strukturtheoretische Ansatz. In M. Rothland (Ed.), *Beruf Lehrer/Lehrerin: Ein Studienbuch* (pp. 103–125). Waxmann.

Hofmann, F., Hagenauer, G., & Martinek, D. (2020). Entwicklung und Struktur der Lehrerinnen- und Lehrerbildung in Österreich. In C. Cramer, J. König, M. Rothland, & S. Blömeke (Eds.), *Handbuch Lehrerinnen- und Lehrerbildung* (pp. 227–236). Klinkhardt.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning* (1st ed.). Erlbaum.

Jesacher-Roessler, L., & Kemethofer, D. (in press). Evidence-informed practice in Austrian education. In C. Brown & J. R. Malin (Eds.), *The Emerald international handbook of evidence-informed practice in education*. Emerald.

Katenbrink, N., & Goldmann, D. (2020). Varianten Forschenden Lernens—Ein konzeptbasierter Typisierungsvorschlag. In M. Basten, C. Mertens, A. Schöning, & E. Wolf (Eds.), *Forschendes Lernen in der Lehrer/innenbildung: Implikationen für Wissenschaft und Praxis* (pp. 195–202). Waxmann.

Kiefer, T., Robitzsch, A., & Wu, M. (2013). *TAM (Test analysis modules).* http://www.edmeasurementsurveys.com/TAM/Tutorials/

Kittel, D., Rollett, W., & Groß Ophoff, J. (2017). Profitieren berufstätige Lehrkräfte durch ein berufsbegleitendes weiterbildendes Studium in ihren Forschungskompetenzen? *Bildung & Erziehung*, *70*(4), 437–452.

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. https://doi.org/10.1016/j.cogdev.2008.09.006

Larcher, S., & Oelkers, J. (2004). Deutsche Lehrerbildung im internationalen Vergleich. In S. Blömke, P. Reinhold, G. Tulodziecki, & J. Wildt (Eds.), *Handbuch Lehrerbildung* (pp. 128–150). Klinkhardt.

Lowman, R. L., & Williams, R. E. (1987). Validity of self-ratings of abilities and competencies. *Journal of Vocational Behavior, 31*(1), 1–13. https://doi.org/10.1016/0001-8791(87)90030-3

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064

Mandinach, E. B., Honey, M., Light, D., & Brunner, C. (2008). A conceptual framework for data-driven decision making. In M. Honey (Ed.), *Data-Driven School Improvement: Linking Data and Learning*, (pp. 13–31). Teachers College Press.

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*(11), 1–48.

Meissner, D., Vogel, E., & Horn, K.-P. (2012). Lehrerausbildung in Baden-Württemberg seit 1945. Angleichungs-und Abgrenzungsprozesse. In C. Cramer (Ed.), *Lehrerausbildung in Baden-Württemberg. Historische Entwicklungslinien und aktuelle Herausforderungen,* (pp. 33–62). IKS Garamond.

Mintrop, R., & Zumpe, E. (2019). Solving real-life problems of practice and education leaders' school improvement mind-set. *American Journal of Education*, *125*(3), 295–344. https://doi.org/10.1086/702733

Münchow, H., Richter, T., von der Mühlen, S., & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network, and relevance for academic success at the university. *The British Journal of Educational Psychology*, *89*(3), 501–523. https://doi.org/10.1111/bjep.12298

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*, *12*(3), 38–50.

Prenger, R., & Schildkamp, K. (2018). Data-based decision making for teacher and student learning: A psychological perspective on the role of the teacher. *Educational Psychology*, *38*(6), 734–752. https://doi.org/10.1080/01443410.2018.1426834

Prenzel, M., Walter, O., & Frey, A. (2007). PISA misst Kompetenzen. *Psychologische Rundschau*, *58*(2), 128–136. https://doi.org/10.1026/0033-3042.58.2.128

Reeves, T. D., & Honig, S. L. (2015). A classroom data literacy intervention for pre-service teachers. *Teaching and Teacher Education*, *50*, 90–101. https://doi.org/10.1016/j.tate.2015.05.007

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. https://doi.org/10.1080/00223891.2010.496477

Richter, D., Böhme, K., Becker, M., Pant, H. A., & Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten. Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. *Zeitschrift für Pädagogik*, *60*(2), 225–244. https://doi.org/10.25656/01:12846

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

Rueß, J., Gess, C., & Deicke, W. (2016). Forschendes Lernen und forschungsbezogene Lehre – empirisch gestützte Systematisierung des Forschungsbezugs hochschulischer Lehre. *Zeitschrift für Hochschulentwicklung, 11*(2), 23–44.

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.

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*(3), 482–496. https://doi.org/10.1016/j.tate.2009.06.007

Schildkamp, K., & Lai, M. K. (2013). Conclusions and a data use framework. In K. Schildkamp, M. Kuin & L. Earl (Eds.), *Data-based decision making in education* (pp. 177–191). Springer.

Schildkamp, K., & Poortman, C. (2015). Factors influencing the functioning of data teams. *Teachers College Record*, *117*(4), 1–42.

Schladitz, S., Groß Ophoff, J., & Wirtz, M. (2015). Konstruktvalidierung eines Tests zur Messung bildungswissenschaftlicher Forschungskompetenz. *Zeitschrift für Pädagogik*, *61*, 167–184. https://doi.org/10.25656/01:15509

Schratz, M., Wiesner, C., Rößler, L., Schildkamp, K., George, A. C., Hofbauer, C., & Pant, H. A. (2018). Möglichkeiten und Grenzen evidenzorientierter Schulentwicklung. *Nationaler Bildungsbericht Österreich*, *2*, 403–454. http://doi.org/10.17888/nbb2018-1.4

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

Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. (2019). The education system in the Federal Republic of Germany 2016/2017. A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe. https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/Bildungswesen-engl-pdfs/dossier_en_ebook.pdf.

Standing Conference. (2020). Ländervereinbarung über die gemeinsame Grundstruktur des Schulwesens und die gesamtstaatliche Verantwortung der Länder in zentralen bildungspolitischen Fragen (Beschluss der Kultusministerkonferenz vom 15.10.2020). https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_10_15-Laendervereinbarung.pdf.

Standing Conference. (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004*. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/ 2004_12_16-Standards-Lehrerbildung.pdf

Stelter, A., & Miethe, I. (2019). Forschungsmethoden im Lehramtsstudium – aktueller Stand und Konsequenzen. *Erziehungswissenschaft*, *30*(58), 25–33. https://doi.org/10.3224/ezw.v30i1.03

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617. https://doi.org/10.1007/BF02294821

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. https://doi.org/10.1177/1094428114553062

Ulrich, I., & Gröschner, A. (2020). Praxissemester im Lehramtsstudium in Deutschland: Wirkungen auf Studierende. Springer VS.

van Geel, M., Keuning, T., Visscher, A. J., & Fox, J.-P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal*, *53*(2), 360–394. https://doi.org/10.3102/0002831216637346

van Geel, M., Keuning, T., Visscher, A., & Fox, J.-P. (2017). Changes in educators' data literacy during a data-based decision making intervention. *Teaching and Teacher Education*, *64*, 187–198. https://doi.org/10.1016/j.tate.2017.02.015

van Ophuysen, S., Behrmann, L., Bloh, B., Homt, M., & Schmidt, J. (2017). Die universitäre Vorbereitung angehender Lehrkräfte auf Forschendes Lernen im schulischen Berufsalltag. *Journal for Educational Research Online, 9*(2), 276–305. https://doi.org/10.25656/01:14952

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

Wessels, I., Rueß, J., Jenßen, L., Gess, C., & Deicke, W. (2018). Beyond cognition: Experts' views on affective-motivational research dispositions in the social sciences. *Frontiers in Psychology, 9*, 1300. https://doi.org/10.3389/fpsyg.2018.01300

Wiesner, C., & Schreiner, C. (2019). Implementation, Transfer, Progression und Transformation: Vom Wandel von Routinen zur Entwicklung von Identität. Von Interventionen zu Innovationen, die bewegen. Bausteine für ein Modell zur Schulentwicklung durch Evidenz (en). *Praxistransfer Schul- und Unterrichtsentwicklung*, 79–140.

Wilson, M., & Scalise, K. (2006). Assessment to improve learning in higher education: The BEAR Assessment System. *Higher Education*, *52*(4), 635–663. https://doi.org/10.1007/s10734-004-7263-y

Wirtz, M. A., & Böcker, M. (2017). Differential item functioning (DIF). In M. A. Wirtz (Ed.), *Dorsch – Lexikon der Psychologie* (18. Aufl.). Hogrefe.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest Version 2: Generalised item response modelling software.* Australian Council for Educational Research.

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. https://doi.org/10.1111/ldrp.12126

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 (Vol. 1). Springer.

Corresponding authors

Jana Groß Ophoff
Institute for Secondary Education, University College of Teacher Education Vorarlberg, Austria
E-mail: jana.grossophoff@ph-vorarlberg.ac.at

Christina Egger
Institute for Didactics, Teaching and School Development, University College of Teacher Education Stefan Zweig Salzburg, Austria
E-mail: christina.egger@phsalzburg.at