Educational Assessment, Evaluation and Accountability (2019) 31:221–249
https://doi.org/10.1007/s11092-019-09296-6

Learning opportunities in teacher education and proficiency levels in general pedagogical knowledge: new insights into the accountability of teacher education programs

Stefan Klemenz[1] · Johannes König[2] · Niclas Schaper[3]

Received: 25 July 2018 / Accepted: 15 April 2019 / Published online: 9 May 2019

Abstract

This paper examines the effects of learning opportunities on the attainment of different proficiency levels in general pedagogical knowledge among student teachers to provide insights into their learning processes and the effectiveness of teacher preparation. The authors used a subsample of the EMW study with two time points, comprising the data of 1451 student teachers from 18 universities and teacher training colleges in Germany and Austria. Findings from logistic regression analyses show that measures of learning opportunities significantly affect the development of general pedagogical knowledge. Whereas instructional quality in seminars and lectures on pedagogy shows effects on achieving the lower levels representing theoretical knowledge components, teaching practice measures related to in-school learning opportunities additionally affects the attainment of the highest proficiency level representing practical knowledge components. Implications for the effectiveness of teacher preparation are discussed.

Keywords Assessment · General pedagogical knowledge · Proficiency levels · Longitudinal study · Opportunities to learn · Teacher education

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11092-019-09296-6) contains supplementary material, which is available to authorized users.

✉ Stefan Klemenz
stefan.klemenz@uni-koeln.de; http://www.hf.uni–koeln.de/36097

Johannes König
johannes.koenig@uni-koeln.de; http://www.hf.uni–koeln.de/30560

Niclas Schaper
niclas.schaper@upb.de

Extended author information available on the last page of the article

1 Introduction

The current discourse on the quality of teacher education focuses on how future teachers acquire professional competence and which role higher education plays in this context (e.g., Cochran-Smith et al. 2012; Blömeke et al. 2014; Hascher 2014; Guerriero 2017). General pedagogical knowledge (GPK) is considered an essential cognitive component of teachers' professional competence (Shulman 1987; Bromme 1992; Baumert and Kunter 2006) and constitutes a highly relevant outcome of initial teacher education programs (Darling-Hammond 2006; König et al. 2017a; Sonmark et al. 2017). A fundamental objective of teacher preparation is therefore supporting pre-service teachers' acquisition of GPK (Blömeke 2011). For the purposes of assessing GPK and examining the effectiveness of teacher education, a precise statement is essential whether pre-service teachers achieve absolute criteria or standards. In this respect, the frequently applied norm-referenced test interpretation of performance based on continuous score scales is regarded inappropriate, since it does not allow to verify the attainment of certain criteria (DeMars et al. 2002; Rost 2004; Hartig et al. 2012).

Proficiency level models address this issue by linking numerical test values to concrete competencies related to the test content which enables criterion-referenced interpretations of performance (Pant et al. 2009; Harsch and Hartig 2011). As a result, persons attaining a certain level can be assumed to have specific competencies available and accordingly whether they reached a criterion or not. Although extensive research has been carried out on teacher professional knowledge, only a few studies (e.g., Mathematics teaching in the twenty-first century; Blömeke et al. 2010; König 2009) modelled proficiency levels. Responding to this, Klemenz and König (2019) generated a proficiency level model for GPK, which is put into focus for the present analyses. It is based on the following two theoretical approaches: cognitive complexity and use of instruction-related terminology. On the one hand, the proficiency model enables substantial statements of what pre-service teachers know and can do, concerning GPK on a given scale level. On the other hand, it offers the opportunity to examine which kinds of learning are significant to achieve certain levels (Sheehan 1997).

The opportunities to learn (OTL) concept is a research approach to investigate specific kinds of learning. By describing and analyzing curricula, it informs about content and process quality of the presented learning opportunities and the effectiveness of teacher education programs (McDonnell 1995; Floden 2015; König et al. 2017b). Findings from different studies provide evidence that OTL—regarded as experiences with an intended learning outcome (Tatto et al. 2008)—are significant factors in the acquisition of knowledge (Schmidt et al. 2011a; Blömeke et al. 2014; König and Klemenz 2015, König et al. 2017a). However, current teacher education research still lacks an accurate and extensive measurement of OTL (Blömeke and Kaiser 2012; König et al. 2017a, 2017b). To counteract this, the EMW study collected extensive and low-inference data on OTL used by student teachers (König et al. 2014a, 2017a). OTL of the two sites of professional learning, namely universities and schools (Flores 2016) are incorporated. The study focuses on instructional quality aspects of higher education courses on general pedagogy as well as pedagogical teaching practice experiences. Both components are highly relevant factors for the acquisition of knowledge (Good and Brophy 2007; Muijs and Reynolds 2011; Baumert and Kunter 2013; Lipowsky et al. 2009; Seidel and Shavelson 2007).

The OTL were linked to student teachers' GPK, which was tested using the TEDS-M instrument (König and Blömeke 2009) and modeled as proficiency levels. Thanks to its longitudinal design, the EMW study allows the examination of proficiency level changes over time. The study draws on a sample of 1451 student teachers from Germany and Austria, who were assessed at two time points. The originality of this study lies in the incorporation of the proficiency level model in our analyses. It makes a major contribution to current research by enabling new insights into which OTL affect the achievement of different levels in GPK and therefore provides valuable information for education policy and teacher education institutions.

1.1 Teacher professional knowledge as an outcome of higher education

Following the expertise research paradigm that takes a cognitive perspective on the teaching profession, an extensive knowledge base is considered crucial for mastering a variety of job-specific tasks and becoming a professional teacher (Bromme 1992; Berliner 2001; Blömeke and Kaiser 2012). Thus, institutions involved in the preparation of teachers are accountable for ensuring the development of pre-service teachers' professional knowledge.

The study's framework draws on a shared conceptualization of professional competence that distinguishes between cognitive and affective-motivational components relevant for successful teaching (Shulman 1987; Weinert 2001; Baumert and Kunter 2006; Shavelson 2010). According to current research, teachers' cognitive competence component can be divided into three different facets of professional knowledge (Shulman 1987; Bromme 1992; Baumert and Kunter 2006): content knowledge (CK), pedagogical content knowledge (PCK), and general pedagogical knowledge (GPK). This multidimensional structure of professional knowledge has been validated for different types of teachers from various countries (König et al. 2011). While CK and PCK are subject-related, GPK is a generic facet that focuses on cross-subject teaching tasks such as classroom management. On a conceptual level, CK and GPK can be distinguished rather clearly (König et al. 2018b). PCK, on the other hand, represents a "special amalgam of content and pedagogy that is uniquely the province of teachers, their own special form of professional understanding" (Shulman 1987, p. 8). Containing aspects of pedagogical and content knowledge, PCK has proven less distinguishable both conceptually and empirically (König et al. 2018b).

Teacher's GPK comprises "those broad principles and strategies of classroom management and organization that appear to transcend subject matter" (Shulman 1987, p. 8) as well as "knowledge about learners and learning, assessment, and educational contexts and purposes" (König 2013, p. 1001). Building on this definition, a review of current empirical studies that assess GPK directly shows agreement that instruction constitutes the core activity of teachers (König 2014). Planning, organization, and reflection of teaching and learning processes as well as their assessment and systemic evaluation are regarded as teachers' core tasks (KMK 2004, p. 3) and therefore as the main focus of GPK. In line with the concept of competence (Weinert 2001; Bromme 2001), the following study considers competencies as latent cognitive dispositions that relate functionally to the mastering of professional tasks (Klieme and Leutner 2006).

Regarding the measurement of competencies, one can distinguish between competence structure and proficiency level models (Hartig and Klieme 2006). Structure models focus on content dimensions and differentiate the type and number of sub-facets of competence. Proficiency level models, on the other hand, qualitatively characterize individuals' abilities on a given level (Hartig and Klieme 2006; Schaper 2009).

1.1.1 Content dimensions of pre-service teachers' general pedagogical knowledge

The present study uses the GPK test instrument developed in the TEDS-M study (König et al. 2011). Drawing on instructional models of effective teaching (Good and Brophy 2007; Helmke 2003; Slavin 1994) and didactics (Good and Brophy 2007; Klafki 1985), four generic dimensions of teaching quality were identified: to prepare, structure, and evaluate lessons ("structure"); to motivate and support student learning as well as to manage the classroom ("motivation/classroom management"); to deal with heterogeneous learning groups in the classroom ("adaptivity"); and to assess students achievement ("assessment"). These content dimensions correspond to the requirements formulated as core tasks in the German national educational standards for teacher education (KMK 2004). In the TEDS-M study, the four dimensions served as a heuristic to develop GPK test items. The test instrument has been subject to multiple validity tests, which provide evidence for construct and curricular validity (see König and Blömeke 2009, König and Seifert 2012, König and Klemenz 2015; Blömeke et al. 2010).

1.1.2 Proficiency levels of pre-service teachers' general pedagogical knowledge

Psychometric models in educational contexts consider test persons' proficiencies usually as continuous latent dimensions (Hartig et al. 2012). Accordingly, the traditional outcome of a test is a continuous score (Sheehan 1997). The key issue is that such scores do not provide information about how proficient a test person is in an absolute sense respective if certain criteria or standards are achieved (Rost 2004; DeMars et al. 2002). Such information, however, is indispensable for decision-making in educational contexts, as "decisions, by definition, create categories" (Cizek 2001, p. 21). At its simplest, whether someone has passed a test or not.

The problem can be solved by proficiency level models which connect test scores to specific cognitive demands. This enables a criterion-referenced interpretation of performance relative to meaningful criteria which are supposed to be essential to achieve (DeMars, Sundre and Wise, 2002). Therefore, proficiency level models account for the requirement to measure adequately teaching-related competences and hence meet emerging licensure needs (ETS 2018). They play an important role in the empirical investigation of student matching of educational standards and for assessing the effectiveness of programs in teacher education (Rost 2004; Harsch and Hartig 2011; Klemenz and König 2019). Furthermore, proficiency models facilitate the communication with different stakeholders about students' achievement (Pant et al. 2009) and enable to investigate cognitive abilities underlying test persons' item responses. "Such additional information may help to better understand the meaning of the kinds of learning which might help to improve those scores" (Sheehan 1997, p. 333).

Nevertheless, modeling proficiency levels are associated with considerable issues. While the division of a continuous measure into categories enables a criterion-referenced description of test scores, it entails a loss of information since the scale is no longer metric. Analyses are hence always less detailed. Proficiency level models play therefore a complementary role in research on professional knowledge and not that of replacing models based on continuous score scales. Furthermore, in particular, the empirical validation of proficiency levels is a central issue (Cizek et al. 2004; Pant et al. 2009). “Critically, standard-setting methods are consensual, normative procedures and there are therefore no innately valid standards or cut scores which could be found or applied” (Tiffin-Richards et al. 2013, p. 15). The result is that although a variety of approaches to model proficiency levels exist (see, e.g., Cizek and Bunch 2007), there is no method that has prevailed “to attain the most valid and defensible interpretations of test-scores” (Tiffin-Richards et al. 2013, p. 15).

In the present study, a repeatedly proved method to model proficiency levels is applied which focuses on the cognitive decomposition of item difficulties (see, e.g., Gorin and Embretson 2006; Hartig 2007; Hartig et al. 2012). Based on theoretical approaches, difficulty-increasing item characteristics are derived and applied to set cut scores by means of statistical procedures (see “Section 2.2.1”). The approach has the advantage of enabling the investigation of construct representation as part of construct validity (Embretson 1983) and is considered an established strategy (Jenßen et al. 2015). “Construct representation concerns the processes, strategies, and knowledge structures that are involved in item solving […]” (Gorin and Embretson 2006, p. 395). If the systematic prediction of item difficulties based on theoretical cognitive approaches succeeds, this underlines the assumptions about the constructs’ cognitive facets (Winther 2010; Hartig and Frey 2012). Thus, reliable information about the validity of test value interpretations is provided (Hartig 2007). How well the cognitive model fits the item difficulties can be quantified with the coefficient of determination (R^2) (Sinharay et al. 2011). In addition, the investigation of construct validity in the context of the nomothetic span by examining relationships of the test score with other relevant variables (Embretson 1983; Messick 1995) can also be performed (see “Section 2.2.1”). Moreover, the approach has the advantage to set cut scores independently from the distribution of test persons’ abilities (Hartig and Frey 2012). Other approaches—e.g., scale anchoring sensu Beaton and Allen (1992) that is for example applied in the PISA studies—divide the score scale at first based on arbitrary criteria (e.g., equal distances on the test value scale) and describe the thresholds post hoc on the basis of appropriate test items near the cut score (Harsch and Hartig 2011).

By modeling proficiency levels, the authors move the focus from the structure of GPK (see “Section 1.1.1”) to cognitive abilities that are essential to solve job-specific tasks. As part of this, the proficiency level model (Klemenz and König 2019) incorporates two characteristics that are highly relevant for meeting teachers’ task requirements: cognitive complexity and the use of instruction-related terminology. Both features refer to core tasks of teaching defined in the German educational standards such as reflecting, evaluating, and organizing teaching and learning processes (KMK 2004; Klemenz and König 2019). Moreover, the characteristics have already proven themselves in a previous study to predict item difficulties (König 2009).

The concept of cognitive complexity refers to approaches proposed by Bieri et al. (1966), Scott (1962), and Peterson and Scott (1983). It focuses on the number of cognitive elements that an individual uses when structuring and assessing certain issues (von Eye 1999). According to this, persons differ regarding the complexity of their cognitive structures in the number of knowledge dimensions they use to solve a task.

With regard to GPK, the authors assume that a high level of cognitive complexity enables multi-perspective views on issues and the creation of differentiated options for action (Klemenz and König 2019). Such abilities support teachers when generating useful performance strategies such as lesson planning and structuring the lesson process, preventing and counteracting classroom disturbances, motivating single students, or the whole group, which are regarded as components of practical knowledge (König 2013). Furthermore, the consideration and weighing of different strategies play a particularly important role in teachers' reflection, which in turn is essential for the development from novice to expert teachers (Berliner 2001).

In addition, the concept of the use of instruction-related terminology is included in the proficiency level model. It can be reasonably assumed that teachers need specific language to access the entire scope of pedagogical knowledge. Taking into account the discussion on teacher knowledge and teacher language (Terhart 1991), three increasingly difficult levels of teachers' instruction-related language were defined: practical, professional, and scientific (König 2009). A profound use of instruction-related terminology at a high language level is expected to indicate a high level of specialized, predominantly theoretical GPK. In such a case, pre-service teachers have in-depth knowledge of key terms and theoretical concepts in the domain of pedagogy and can retrieve them in certain situations, which in turn makes an application in teaching processes more likely (Bromme 2001). Furthermore, it demonstrates the ability to explicate pedagogical knowledge. In particular, reflection on action (Schön 1983) requires knowledge that is explicit to be analyzable and re-organizable (Altrichter and Posch 2007).

1.2 Opportunities to learn in teacher education

According to model-based descriptions of teacher education programs, a fundamental objective of teacher preparation is supporting pre-service teachers' acquisition of professional competence (Blömeke 2011). In order to comply with the requirements (Tatto et al. 2008; Blömeke et al. 2014), institutions and policymakers involved in teacher education create and orchestrate different OTL as "a set of experiences and content exposures" (Schmidt et al. 2008, p. 736). OTL serve as indicators of curricular variation (Tatto et al. 2008) and allow to examine whether differences in learning opportunities are related to differences in professional knowledge (McDonnell 1995). The concepts' starting point is the curriculum, which is considered the most fundamental underlying structure of educational processes (Houang and Schmidt 2008). OTL relate to the distinction between three curriculum representations: intended, implemented, and attained curriculum (McDonnell 1995; van den Akker 2003). The intended curriculum includes the vision of the curriculum as well as the resulting specifications in formal documents (written curriculum), i.e., study and examination regulations (McKenney et al. 2006; Vanderlinde et al. 2009; König et al. 2017a, Tachtsoglou and König 2017). The implemented curriculum refers to different learning activities aiming at the achievement of the intended learning goals. It comprises the "operational curriculum, i.e., the actual process of teaching and learning" (Vanderlinde et al. 2009, p. 574). The attained curriculum results from the intended and implemented the curriculum.

Despite a large variation in the precise design of programs in teacher education, there is consensus that essential components exist: subject knowledge, subject didactic knowledge, general education studies, and practicum (e.g., Flores 2016; Kansanen 2014; Schmidt et al. 2011a). Considering these key components, the "two sites of professional learning (schools and universities)" (Flores 2016, p. 205) during initial teacher education become apparent. On the one hand, universities provide content-related OTL to achieve professional, primarily theoretical knowledge of CK, PCK, and GPK. On the other hand, pre-service teachers are offered in-school OTL to gain experience in teaching practice.

Although empirical studies investigating pre-service teachers OTL increased particularly in the past years, the focus was predominantly on the content students were exposed to during their studies. Various findings provide evidence that content-related OTL conveyed in higher education courses affect cognitive outcomes (e.g., Blömeke et al. 2012; Schmidt et al. 2011b; König et al. 2017a). However, when discussing educational opportunities in teacher education, instructional facets such as teaching quality or teaching methods are highly valuable to take into account (Houang and Schmidt 2008). From school research, it is known that quality aspects of instruction are relevant for effective teaching (e.g., Seidel and Shavelson 2007; Baumert and Kunter 2013). In the context of the TIMSS video study, three basic dimensions of instructional quality were identified: student-oriented climate, classroom management/structuring, and cognitive activation (Klieme et al. 2001). Whereas the first dimension affects primarily students' motivational outcomes, the other two basis dimensions show effects on cognitive achievement (Klieme et al. 2001). These findings are supported by results from the Pythagoras study which determined the impact from both cognitive-oriented OTL on students' mathematics competencies (Rakoczy et al. 2007). Meta-studies report significant effects from cognitive activation and structured teaching on cognitive student outcomes as well, but effect sizes are relatively low (structured teaching: $0.04 < \text{Cohen's } d < 0.22$; cognitive activation: $0.02 < \text{Cohen's } d < 0.06$) (see Seidel and Shavelson 2007; Scheerens and Bosker 1997). However, this does not mean that they are not essential. "This is especially the case in the study of teaching and learning, given the large number of factors affecting students' performance in school systems. Moreover, even a small effect has consequences for thousands of students" (Seidel and Shavelson 2007, p. 471). Although the findings regarding instructional quality related to OTL are obtained in the field of school research, they are expected to be relevant for the acquisition of knowledge in higher education as well. Against this background, the present study focuses on cognitive activation (Baumert and Kunter 2013; Lipowsky et al. 2009; Seidel and Shavelson 2007) and on the structuring of learning processes (Good and Brophy 2007; Muijs and Reynolds 2011) in higher education courses. With regard to the type of knowledge, it is assumed that higher education courses in particularly support the acquisition of theoretical-formal knowledge (Fenstermacher 1994) such as objective facts, theoretical approaches, and central concepts (König 2013, Tachtsoglou and König 2017).

Besides learning opportunities in higher education courses, in-school OTL which primarily aim at giving students opportunities to teach are another relevant component (Tachtsoglou and König 2017). Recent empirical studies found evidence that teaching practice affects teachers' acquisition of professional knowledge (König and Klemenz 2015, König et al. 2017a; Blömeke 2011; Blömeke et al. 2012; Schmidt et al. 2011a). Regarding the development and organization of professional knowledge, teaching practice activities are assumed not only to affect the amount of knowledge positively but also to foster the structuring and proceduralization of knowledge (Voss et al. 2015). Therefore, teaching-related in-school OTL are considered to support the acquisition of teachers' practical knowledge (Kolbe and Combe 2004; König 2013). Not surprisingly, experts in the teacher education community stated that future teachers needed more effective OTL to teach (Lampert and Ball 1999; Cochran-Smith and Villegas 2016). Building on previous studies (e.g., König et al. 2014, 2017a; Tatto et al. 2012), the measures implemented in the EMW study covered four areas of teaching practice: lesson planning, teaching, linking theories to situations, and reflecting on practice. Previous analyses using these teaching content scales indicate that they can substantially explain cognitive outcomes ($0.06 < \beta < 0.18$) (König and Klemenz 2015, König et al. 2017a).

In summary, both components of professional learning—in universities and in schools—contribute to the acquisition of knowledge. Although theoretical knowledge is presumably at least partly acquired in practical OTL and practical knowledge at least partly in content-related OTL, it can be assumed that certain learning opportunities are more appropriate to support certain types of knowledge. Following Berliner (2004), theoretical and practical knowledge are highly relevant types of knowledge and play an essential role in the progress from the stage of novices to advanced beginners. Whereas central terms and concepts as well as context-independent rules of teaching are the subject matter of novices' learning (König 2010) and form a solid knowledge base, the advanced beginner starts with the acquisition of practical knowledge which is contextually developed and considered as knowledge based on experiences made during teaching practice. Against this background, a central aim of teacher education systems is to support future teachers in their task to acquire and connect theoretical as well as practical issues of teaching (Clift and Brady 2005) as both types of knowledge contribute to the expert teachers' performance in the classroom (Bromme 2001). The significance to provide certain OTL by teacher education institutes in order to develop professional knowledge components can be outlined using the example of acquiring classroom management expertise, which is considered as one of the most essential domains in the teaching profession and part of teacher GPK (e.g., Hattie 2009; Kunter et al. 2011; Seidel and Shavelson 2007). On the one hand, theoretical knowledge about classroom management (e.g., strategies to prevent and counteract interferences, effective use of allocated time, and routines) is essential and can be well conveyed in seminars and lectures. On the other hand, this theoretical knowledge needs to be proceduralized (Anderson 1982), embedded in contexts, and integrated with practical knowledge in the field (Clift and Brady 2005), which can be supported through teaching-related in-school OTL. It is considered substantial that both components of OTL are provided and linked to ensure sufficient congruence between "what is taught at the university and what students experience during their field experiences" (Jones 2006, p. 895).

Regarding the measurement of learning opportunities, critical voices have pointed out that quantitative studies use primarily distal and aggregated indicators such as degrees, type of license, or the number of courses taken to define OTL (Blömeke et al. 2014; Schmidt et al. 2011b). To meet the demand of less aggregated measures of OTL, pre-service teachers had been exposed to in their teacher preparation programs, a large number of items and scales were generated to allow for a detailed analysis of OTL (see "Section 2" for more details).

1.3 Teacher education in Germany and Austria

Our analysis is based on data from the two German-speaking countries Germany and Austria which have a similar linguistic and cultural background (see König and

Klemenz 2015). Furthermore, students from both countries participate in teacher preparation programs in universities or teacher training colleges which provide courses on general pedagogy as well as in-school OTL supporting teaching practice. However, there are also some, particularly structural differences regarding the concrete design and provision of learning opportunities. Such variations in OTL and potential learning outcomes serve as good reasons for the examination. Moreover, features of the focused programs are similar to features of teacher preparation programs more broadly (König et al. 2017a). This leads to results that can be generalized and contribute to our understanding of learning processes in teacher education.

The national educational standards for teachers in Germany (KMK 2004) provide an orientation framework for the design of curricula at German universities. The curricula in Austria are also based on German educational standards and hence show similarities to those in Germany. Both countries emphasize the study of general pedagogy which aims at supporting the development of different knowledge facets and evidence-based reflection about teaching and learning. The component of general pedagogy comprises various facets such as didactics, educational psychology, theories of schooling, methods of educational research, assessment, and teaching methods (Tatto et al. 2008). Concerning teaching practice, the bachelor program in Germany is designed with only a few weeks of practical elements. This resulted from the split into a first phase at the university with a strong emphasis on theory and a subsequent second practical phase (induction phase) in which future teachers teach regularly at school. By contrast, most teachers in Austria receive teaching practice right from the start when they enter initial teacher education and to a much greater extent. In particular, teacher training colleges which do not differentiate between two phases due to the comparatively short initial training of teachers (3 years at the time point of data collection)[1] offer teacher training courses in which theoretical learning is closely linked to teaching practice in schools. This is also reflected in findings from an analysis comparing teacher education programs from both countries showing that the number of practical units is significantly higher in Austria than in Germany (Arnold 2014).

1.4 Research questions

The present study aims to address the following research questions and assumptions:

1. Do learning opportunity measures of instructional quality of lectures and seminars as well as facets of teaching practice serve as predictors for the achievement of higher proficiency levels in GPK among student teachers?

H1: Taking into account that the individual use of OTL by student teachers has a positive influence on the acquisition of knowledge (see “Section 1.2”), the authors assume that the attainment of more advanced proficiency levels in GPK should be substantially predicted by indicators measuring instructional quality and the amount of teaching practice activities. The effects should be significant even when controlled for

[1] Teacher education in Austria has undergone a far-reaching reform in 2016/2017 that provides for an expansion of academic components during teacher preparation for students of all types of teaching careers. For example, primary school teachers have to study a 4-year bachelor today.

student teachers' background variables, entry characteristics, and learning opportunities related to pedagogical content.

2. Do facets of instructional quality of higher education courses have a substantial effect particularly on attaining lower proficiency levels among student teachers?

H2: Lectures and seminars in higher education are suitable for conveying student teachers' to central terms and concepts, which are fundamental at the beginner stage of teacher professional development (see "Section 1.2"). Thus, it can be assumed that this kind of OTL primarily supports the acquisition of theoretical-formal pedagogical knowledge. As the lower proficiency levels are characterized by language levels and the associated theoretical knowledge (see "Section 2.2.1"), we assume, therefore, that quality aspects of instruction in higher education courses on general pedagogy should have a particularly significant impact on achieving the lower levels of the model.

3. Do facets of teaching practice have a substantial effect on attaining the highest proficiency level among student teachers?

H3: Student teachers who attain the highest proficiency level are capable of solving tasks, which demand complex cognitive processes (see "Section 2.2.1"). They are able to take multiple perspectives and to create a variety of options and strategies for teacher action (Klemenz and König 2019). In line with findings that teaching practice activities support the acquisition of practical knowledge (see "Section 1.2"), we assume that especially these activities support the ability to master complex cognitive, instruction-related tasks. Therefore, we hypothesize that teaching practice significantly affects the achievement of the highest level of competence.

2 Method

2.1 Participants and procedures

This study uses data from the EMW study ([Entwicklung von berufsspezifischer Motivation und pädagogischem Wissen in der Lehrerausbildung] Change of Teaching Motivations and Acquisition of Pedagogical Knowledge during Initial Teacher Education), funded by the Rhine-Energy-Foundation Cologne, Project number W-13- 2-003 and W-15-2-003), an empirical study on teacher preparation at higher education institutions in the three German-speaking countries Germany, Austria, and Switzerland. The longitudinal design of the EMW study enables the capturing of student teachers' GPK development over time. Supported by a network of about 40 research partners from Germany, Austria, and Switzerland in autumn 2011, 6601 student teachers from 31 universities and teacher training colleges were sampled in their first semester, representing a population of nearly 50,000 student teachers at the beginning of their bachelor studies (see König et al. 2017a). The survey was primarily conducted in large lectures where students' attendance was compulsory in order to limit individual self-selection bias. Since the recruitment of the respondents at the second time point proved to be not feasible at some institutions, few had to be excluded. A response rate at the second measurement point of 56% in Germany and 73% in Austria was achieved. To answer the research questions presented, a sub-dataset of the EMW study is used, which includes teacher training students from Germany and Austria, who could be followed up to participate at both time points in autumn 2011 (1st semester) and autumn 2013 (5th semester). The subset consists of 1451 bachelor student teachers from 18 universities and teacher training colleges (see Table 1). At the first occasion of measurement, the participants had recently started their bachelor studies. At the second time point, they were in the last year of their bachelor studies. Student teachers from four different types of teaching careers participated (primary school teaching, lower secondary school teaching (Haupt-/Real-/Gesamtschule), lower and upper secondary school teaching (Gymnasium/Gesamtschule), and special needs education). Thus, the subset covers a broad range of program types.

Due to the aforementioned sample failure, drop-out analyses were carried out separately for the German and Austrian sample to investigate whether a selection bias exists. Based on the sample of the first occasion the participation in the second occasion was predicted using binary regression analyses. Included were the background variables age and gender, the performance variables grade and score in GPK, and eleven scales from the FIT-Choice instrument measuring motivations for selecting teaching as a career (see Watt and Richardson 2007). Regarding the German sample, four motivational scales are significant (job security, time for family, work with children/adolescents, and social influences), but effect sizes are very small (0.005 < average marginal effects < 0.012). Regarding the Austrian sample, three motivational scales (fallback career, work with children/adolescents, and social influences) and the GPK test score are significant. Effect sizes are very small as well (0.005 < average marginal effects < 0.011). The results of the drop-out analysis therefore do not point to a strong selection bias. Nevertheless, as small effects exist, the results should be interpreted with some caution.

2.2 Assessment of general pedagogical knowledge

Student teachers' GPK was assessed via a standardized paper-and-pencil test developed in the TEDS-M study (König and Blömeke 2009) capturing the four generic teaching dimensions presented above (adaptivity, structure, classroom management/motivation,

Table 1 Pre-service teacher and institution samples by country and type of teacher training

	Germany		Austria		Total	
Type of teacher training	Pre-service teachers	Institutions	Pre-service teachers	Institutions	Pre-service teachers	Institutions
Primary school	208	5	386	6	594	11
Lower secondary school	239	5	194	5	433	10
Lower and upper secondary school	136	5	46	1	182	6
Special needs education	207	4	35	2	242	6
Total	790	10	661	8	1451	18

assessment). For each dimension of GPK, a subset of items was designed (for an item example, see Table 2). The test instrument includes open-response as well as multiple-choice items, which are equally distributed across the four teacher tasks. The coding of the open-response answers was carried out by trained raters using the comprehensive coding manual from the TEDS-M study. The interrater agreement showed good results (Cohen's Kappa $M = 0.80$).

Using the software ConQuest (Wu et al. 1997), GPK test data were IRT scaled following a procedure that has proven powerful in previous studies (König and Seifert 2012, König and Klemenz 2015). Initially, the authors analyzed the test data in the one-dimensional Rasch model from each of the two occasions of measurement separately to examine the invariance of the two item parameter sets deriving from each scaling analysis.

$$P(X_{is} = 1|\theta, \beta_i) = \frac{e^{(\theta_s-\beta_i)}}{1+e^{(\theta_s-\beta_i)}}$$

The reliability of both scaling procedures was acceptable (first occasion: EAP/PV reliability 0.73, WLE reliability 0.68; second occasion: EAP/PV reliability 0.70, WLE reliability: 0.69; both comparable to Cronbach's alpha). In the next step, item difficulties from both occasions of measurement were correlated. The high correlation ($r = .79$) indicates sufficient independence of the samples so that a concurrent scaling can be considered permissible (see Bond and Fox 2007). Subsequently, the authors treated all observations as independent and carried out one-dimensional Rasch scaling with all observations from both measurement occasions (concurrent scaling) to increase the analytical power of our final scaling analysis (Bond and Fox 2007; Rost 2004). The concurrent calibration was conducted with complete equality of all item parameters for both groups and without fixed parameters (see Von Davier and von Davier 2007). The reliability of the concurrent scaling was good (EAP/PV reliability 0.80, WLE reliability 0.77). Item difficulty values spread over a range of around five logits (−2.78 to 2.20); the item discrimination values have an average of .35 (min 0.10, max 0.65). The weighted mean squares fall within an acceptable range of 0.80 to 1.20. Only one item has a value of 1.21. For theoretical reasons, the item was not excluded from the analysis.

Table 2 Item example from the TEDS-M GPK test, dimension structuring

Imagine you are helping a future teacher to evaluate her lesson because she has never done this before. To help her adequately analyze her lesson, what question would you ask?

1)
2)
3)
…
10)

Formulate ten essential questions and write them down.

2.2.1 Modeling proficiency levels

To model proficiency levels, an established procedure predicting item difficulties by item attributes was applied (see, e.g., Gorin and Embretson 2006; Hartig 2007; Hartig et al. 2012). The approach can be divided up into three stages.

First, item attributes are derived from the theoretical approaches (see "Section 1.1.2") and each test item is assigned to a specific level of the item attributes (Gorin and Embretson 2006; Hartig 2007; Hartig et al. 2012). Based on the concept of cognitive complexity, the item attribute complexity of cognitive processes was derived (Klemenz and König 2019). It is assumed that test items of higher complexity call for broader cognitive activity and impose respectively higher demands on the test person (Klemenz and König 2019; Kauertz et al. 2010). Tasks that require only one knowledge element were assigned to a low level of cognitive complexity, tasks requiring several, connected knowledge elements to a high level. The amount of required knowledge elements is defined by the tasks' coding scheme as determined in the comprehensive coding manual (see König et al. 2011). For example, the item in Table 2 was classified as complex because several linked knowledge elements respective items are required for the solution. If a test person solves such a complex task, this is assumed as an indicator for higher cognitive complexity as outlined above (see "Section 1.1.2"). Tasks that require only one item and therefore do not call for complex cognitive processes were classified to a low level. As a result of this approach, 14 items were classified to the high level and 34 to the low level of complexity of cognitive processes (see Table 3).

Building on the concept of domain-specific use of terminology, i.e., terminology relevant for teaching in general, the item attribute terminological requirements were derived. Initially, the test item stems were analyzed rather than the tasks responses to determine which instruction-related terms they contained. Thus, 86 terms were identified, which were then assigned to the three language levels (see "Section 1.2") by eight teachers, educationalists, and psychologists, all working in teacher education (König 2009). The eight independent ratings show very good reliability ($\text{ICC}_{\text{unjusted}}$ (2, 8) = 0.95). Based on the averages of all eight ratings, 18 items were assigned to the practical (e.g., "group work," "pace of instruction"), 24 items to the professional (e.g., "differentiated learning," "cognitive learning goal"), and 6 items to the scientific (e.g., "reliability," "operant conditioning") language level. The test item in Table 2 was assigned by the eight raters to the practical language level. The test instrument item pool does not contain items that combine complex cognitive processes with a scientific language level so that the model cannot represent a level with the most sophisticated combination of features. Table 3 shows the distribution of test items by a combination of item attributes.

In a second step, item difficulties obtained from IRT-scaling are predicted by item attributes using multiple regression analysis. The coefficient of determination (R^2) provides "an idea of how well the item statistics can be predicted by the item attributes" (Sinharay et al. 2011, p. 65) and hence if scale anchoring provides useful information. With 43%, a large part of the variance between item difficulties can be explained by the underlying cognitive model ($R_{\text{adjusted}} = .43$) (Klemenz and König 2019). The amount of explained variance is comparable to other studies applying the same approach (Hartig et al. 2012) and supports the validity of test interpretations with regard to construct representation.

Table 3 Distribution of test items by a combination of item attributes

Complexity	Language level	Number of items
Simple	Practical	9
Simple	Professional	19
Simple	Scientific	6
Complex	Practical	9
Complex	Professional	5
		48

Third, based on the results of the regression analysis cutoff points of the proficiency levels can be determined using an additive approach. The cut scores result from the addition of the constant with the unstandardized regression coefficient of each item attribute (Hartig 2007; Klemenz and König 2019). Four proficiency levels in GPK were generated (see Table 4).

Due to the cognitive understanding of competence in this study, an analysis appears necessary whether intelligence plays a role in the measurement and thus whether the construct could be at least partially confounded (Messick 1995). Since the construction of proficiency levels is not based on general cognitive skills, but rather on cognitive abilities in GPK, the differences in school grades—as an indicator of general cognitive skills—between test persons on higher and lower levels are assumed to be small. Although the mean differences of the grades are significant across the four levels, the effect sizes are very small (Germany: $F(3, 1491) = 5.247$, $p < 0.01$, $\eta^2 = 0.01$; Austria: $F(3, 1218) = 7.899$, $p < 0.01$, $\eta^2 = 0.02$). The results can be interpreted as an indication of discriminatory validity (König and Seifert 2012).

2.3 Surveying opportunities to learn

In order to comply with the aspiration to measure OTL in a low-inference way, the authors developed a large range of items covering key components of instructional quality of lectures and seminars and in-school opportunities for teaching practice. The instruments were designed and implemented within the framework of the EMW study

Table 4 Proficiency levels with short description and thresholds (logits) (Klemenz and König 2019)

Proficiency level	Logit	Short description
Below I	< − 0.942	Test persons cannot solve simple cognitive processes on a practical language level with sufficient probability.
I	− 0.942	Test persons can solve simple cognitive processes on a practical language level with sufficient probability.
II	− 0.374	Test persons can solve simple cognitive processes on a professional/scientific language level with sufficient probability.
III	0.989	Test persons can solve complex cognitive processes on a practical/professional language level with sufficient probability.

(see König et al. 2014, 2017a). The instrument for measuring the instructional quality of courses comprises 29 items, which are classified into four scales: structured teaching (lectures/seminars) and cognitive activation (lectures/seminars). Table 5 provides item examples for each sub-area and scale reliabilities. The distinction between seminars and lectures allows a detailed analysis of teaching formats. Student teachers' responses were prompted by the request: "In the lectures/seminars I have attended so far on pedagogical topics and contents…" The response categories ranged from "does not apply at all" to "fully applies" (4-point Likert scale). The scales therefore covered the quality of the four aforementioned components valued by student teachers. Our conceptualization of teaching practice distinguishes between four relevant dimensions: lesson planning, teaching, linking theories to situations, and reflecting on practice (König et al. 2014, 2017a). The instrument consists of 65 items (see Table 5). All items have a dichotomous response format (0 = no/1 = yes) and were introduced in the questionnaire with the following question: "During your teaching practice up to now have you carried out the following activities?" The student teachers indicated whether they had conducted the activities or not.

Table 5 Item examples from the learning opportunities instructional quality of courses, teaching practice, and pedagogical content, number of items, and reliability

Area	Scale	Number of items	Item example	α
Instructional quality of courses	Structured teaching (lectures)	3	…the knowledge was conveyed in a well-structured manner.	0.761
	Cognitive activation (lectures)	5	…the students dealt with pedagogical questions in an intellectually demanding way.	0.823
	Structured teaching (seminars)	3	…the contents were clearly presented.	0.819
	Cognitive activation (seminars)	7	…the students took an active part in plenary discussions.	0.857
Teaching practice	Lesson planning	12	I have formulated learning goals aligned with the curriculum.	0.822
	Teaching	31	I have checked attendance.	0.897
	Linking theories to situations	11	I have observed teaching methods that I have learned at my university/teacher training college course.	0.826
	Reflecting on practice	11	I have drawn conclusions for future teaching.	0.776
Content	Pedagogical content	37	Differentiated instruction Analyzing own teaching Whole-class motivation.	0.889

Note: α—Cronbach's Alpha

(a) Introductory question (Response format): In the lectures/seminars I have attended so far on pedagogical topics and contents…? ("does not apply at all" to "fully applies," 4-point Likert scale)

(b) Introductory question (Response Format): During your teaching practice up to now have you carried out the following activities? (yes/no)

(c) Introductory question (Response Format): Have you ever studied the relevant topic aspect? (yes/no)

In addition, pedagogical content topics pre-service-teachers had studied were measured by a total of 37 items comprising four sub-areas (adaptivity, structure, classroom management/motivation, and assessment) corresponding to the design of the GPK test (see Table 5). The items require to indicate whether or not they have ever studied the relevant topic aspect, and they have to answer with "yes" (coded as 1) or "no" (coded as 0). As the focus of this study is on quality facets of courses and teaching practice, we incorporated merely a total score of the content dimensions into the analyses as control variable.

From Table 5, it is apparent that Cronbach's alpha as a measure of internal consistency indicates good scale reliabilities for all of the OTL scales (0.761 < Cronbach's alpha < 0.897). The intercorrelations between the scales of each OTL facet are low to moderate (quality scales 0.196 < Pearson's $r < 0.565$; teaching practice scales 0.542 < Pearson's $r < 0.675$; all $p < 0.001$) (see Tables 6 and 7). For ease of interpretation, we chose to consider each scale separately in the analyses.

2.4 Missing values

The sub-dataset of 1451 respondents contained 95.49% complete data. Prior to our regression analyses, the authors cautiously handled the missing data via multiple imputation method and determined that the missing data was missing at random (MAR). For missing data, the authors used multivariate imputation by chained equations in R (Package "mice," version 2.46.0) with predictive mean matching (PMM) and 10 iterations to complete the dataset (Manly and Wells 2015; van Buuren and Groothuis-Oudshoorn 2011). The imputation model used all the variables that the authors considered potential predictors. In total, 1244 (4.51%) of the required 27,569 values were imputed.

2.5 Background and entry characteristics

To account for individual characteristics of student teachers, the authors controlled for demographics (age, gender), and the social status by including the highest socio-economic index of both parent in our analyses (HISEI, following the concept of International Socio-Economic Index of Occupational Status by Ganzeboom et al. 1992). Moreover, it was accounted for academic performance by including the secondary school grade point average and adjusted for the GPK scores from the first time point. Furthermore, the authors accounted for the country (Germany vs. Austria).

Table 6 Matrix of the bivariate manifest correlations of quality of instructions scales

Scale	1	2	3	4
1. Structured teaching (lectures)	1			
2. Cognitive activation (lectures)	0.364**	1		
3. Structured teaching (seminars)	0.565**	0.196**	1	
4. Cognitive activation (seminars)	0.261**	0.373**	0.540**	1

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

Table 7 Matrix of the bivariate manifest correlations of teaching practice scales

Scale	1	2	3	4
1. Lesson planning	1			
2. Teaching	0.660**	1		
3. Linking theories to situations	0.616**	0.542**	1	
4. Reflecting on practice	0.675**	0.581**	0.653**	1

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

3 Results

3.1 Descriptive statistics

Figure 1 shows the frequency distribution of student teachers over the proficiency levels at each occasion of measurement. The results agree well with the assumption that knowledge is acquired during higher education studies: in the advanced stage of their studies (2nd time point, 5th semester bachelor's program), student teachers achieve higher proficiency levels significantly more often than at the beginning of their studies (1st time point, 1st semester bachelor's program) ($\chi^2 = 763.33$, df = 3, $p < 0.01$).

Comparing the means on the OTL scales (see Table 8), student teachers report instructional quality values slightly above or even slightly below ("cognitive activation (lectures)") each of the scale mid-points. The mean differences between German and Austrian pre-service teachers are significant except for the scale "cognitive activation (seminars)". However, effect sizes are small ($0.001 < \eta^2 < 0.035$). The ICCs of the quality scales demonstrate that only the scale "cognitive activation (lectures)" shows considerable variation between different programs ($n = 37$) defined as the teacher preparation program within the single university or training college. Regarding the teaching practice that student teachers made use of, results present a different picture as all mean differences are significant and effect sizes range from middle to large ($0.091 < \eta^2 < 0.259$). Furthermore, the ICCs demonstrate that large variance exists between the institutions as well ($0.261 < \text{ICC} < 0.392$). The results concerning the content scale however indicate that there is relevant variance between institutions (ICC = 0.332) but not between countries ($\eta^2 = 0.001$).

3.2 Findings from ordinal regression analysis

To answer the first research question (H1), it was analyzed if each of the OTL facets has a significant effect on achieving higher proficiency levels. Eight ordinal logistic regression models (one OTL scale each; proportional odds models (POM)) were carried out to predict the dependent, ordinal variable GPK (four proficiency levels) at the second measurement point using the software R (package "ordinal," version 2015.6-28 by Christensen 2015) (see Tables 9 and 10). POM assume equal slopes (proportional odds) across all levels (Agresti 2010). The equal slope assumption is in none of the following models violated. Before we

[0] Nonetheless, to confirm our findings, multi-level logistic regressions (ordinal and binary, see "Section 3.3") were conducted. Overall, however, as based on the low ICCs expected the results obtained do not lead to different interpretations. Findings from multi-level regressions are nearly identical.

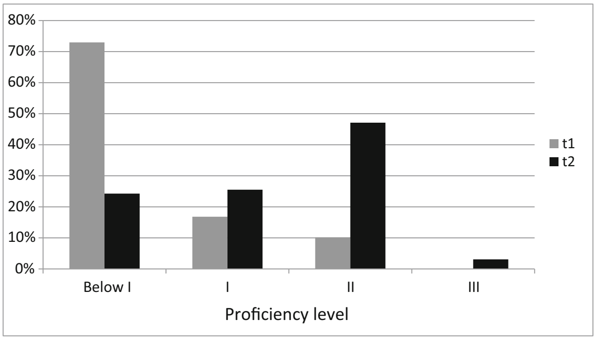

Fig. 1 Frequency distribution of the pre-service teachers on proficiency levels at 1st and 2nd time point (Klemenz and König 2019)

analyzed the models with all predictors included, an empty model was fitted first, in order to assess the dependencies at the program level. To assess the dependencies due to the students being nested within the same programs, intra-class correlations (ICCs) were computed. As the ICC of the empty model was low (ICC = 0.058) and therefore the application of multi-level models is not crucial, single regression models are carried out in the following analyses.[2] This decision was moreover based on the relatively small group size on the highest level which led to unstable parameter estimations, in particular concerning the binary regression analyses on the highest proficiency level (see "Section 3.3").

In addition to each scale, the authors entered the control variables country (Germany = 0, Austria = 1), age, gender, school grade, the highest ISEI (HISEI), the proficiency level achieved at the first measurement point (as dummy variables), and the OTL content scale into the regression equation. Moreover, the authors *z*-standardized

Table 8 Descriptive statistics of OTL scales (mean, standard deviation (SD), standard error (SE), η^2, ICC, minimum (Min), and maximum (Max))

Scale	Germany		Austria		*p* value	η^2 (countries)	ICC (programs)	Min/Max
	Mean	SD/SE	Mean	SD/SE				
Structured teaching (lectures)	2.719	0.534/0.020	2.632	0.516/0.021	0.002	0.007	0.059	1/4
Cognitive activation (lectures)	2.164	0.537/0.020	2.380	0.593/0.024	0.001	0.035	0.137	1/4
Structured teaching (seminars)	2.835	0.516/0.019	2.751	0.575/0.023	0.005	0.006	0.044	0/4
Cognitive activation (seminars)	2.718	0.506/0.019	2.689	0.537/0.021	0.309	0.001	0.031	0/4
Lesson planning	5.241	2.962/0.110	8.453	2.375/0.095	0.001	0.259	0.392	0/12
Teaching	17.716	6.470/0.240	21.952	6.897/0.277	0.001	0.091	0.363	0/31
Linking theories to situations	5.213	2.976/0.113	8.033	2.292/0.096	0.001	0.215	0.304	0/11
Reflecting on practice	4.469	2.712/0.101	6.940	2.451/0.100	0.001	0.184	0.261	0/11
Pedagogical content	23.391	7.525/0.275	23.285	7.601/0.303	0.522	0.001	0.332	0/37

Table 9 Ordinal regression analysis: GPK at the second occasion containing the quality scales with parameter estimates (regression coefficient B)

Predictor	M1 B	M2 B	M3 B	M4 B
Country[a]	− 0.508***	− 0.512***	− 0.489***	− 0.496***
Age	0.029	0.031	0.034	0.024
Gender	− 0.002	− 0.018	0.002	0.022
Grade	− 0.346***	− 0.34***	− 0.343***	− 0.34***
HISEI	0.022	0.021	0.026	0.021
GPK 1, level I	0.526***	0.522***	0.525***	0.52***
GPK 1, level II	1.031***	1.021***	1.027***	1.024***
OTL Content	0.267***	0.291***	0.251***	0.253***
OTL structured teaching (lectures)	0.037			
OTL cognitive activation (lectures)		− 0.068		
OTL structured teaching (seminars)			0.122*	
OTL cognitive activation (seminars)				0.158**
Nagelkerke R^2	0.092	0.093	0.096	0.098

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

[a] Germany = 1, Austria = 2

continuous entry characteristics, background variables, and instruments (age, grade, HISEI, and OTL scales) before incorporating them into the regression equation. With regard to the facets of instructional quality of lectures and seminars (Table 9), only the

Table 10 Ordinal regression analysis: GPK at the second occasion containing the teaching practice scales with parameter estimates (regression coefficient B)

Predictor	M5 B	M6 B	M7 B	M8 B
Country[a]	− 0.588***	− 0.585***	− 0.624***	− 0.578***
Age	0.028	0.023	0.029	0.028
Gender	− 0.012	0.014	− .013	− 0.011
Grade	− 0.348***	− 0.352***	− .357***	− 0.35***
HISEI	0.02	0.021	0.031	0.023
GPK 1, level I	0.519***	0.532***	0.532***	0.516***
GPK 1, level II	1.02***	1.011***	1.049***	1.022***
OTL content	0.251***	.218***	0.2**	0.244***
OTL planning	0.083			
OTL teaching		0.178**		
OTL linking			0.173**	
OTL reflecting				0.092
Nagelkerke R^2	0.093	0.099	0.097	0.093

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

[a] Germany = 1, Austria = 2

scales “structured teaching (seminars)” and “cognitive activation (seminars)” show significant effects across all levels. Regarding in-school OTL, only the “teaching” and “linking” scales exhibit significant effects (Table 10). Student teachers who made use of more of these OTL are more likely to attain higher levels of proficiencies. The content scale is significant in all POM models.

3.3 Findings from binary logistic regression analysis

To answer the second and third research questions, binary logistic regression models were conducted. The effects of OTL scales were assessed separately for each of the three thresholds of the proficiency level model. The same control variables as in the case of ordinal regression analyses and *z*-standardized continuous variables were accounted for. In order to enable a comparison of the regression coefficients both within a model and between the models, average marginal effects (AMEs) were calculated. AMEs are in case of binary logistic model comparisons clearly preferable to unstandardized coefficients or odds ratios (Best and Wolf 2012). They indicate an average effect and can be interpreted as follows: if x increases by one unit, the likelihood of $y = 1$ increases on average by AME points (see Best and Wolf 2012). Thus, they can be interpreted as effect sizes in contrast to unstandardized coefficients or odds ratios. The p values of AMEs are Bonferroni-Holm corrected to counteract the issue of multiple comparisons. In order to provide a condensed presentation of the findings of the binary logistic regression analyses, the control variables in the presented visualization form are excluded (see Tables 8 and 9; for details of binary regression analyses, please see Table B1-B6 of the Electronic Supplementary Material). Therefore, the tables represent only the AMEs of the certain OTL scales and include only models with significant effects on at least one of the proficiency levels. As the ICCs of the empty binary models were low ($0.051 < \text{ICC} < 0.099$), multilevel models were dispensed as well.

Table 11 shows the results of the analyses for the scales “structured teaching (seminars)” and “cognitive activation (seminars).” As expected, both scales have a significant impact only on the attainment of the first and second proficiency levels. For

Table 11 Findings from binary logistic regression analyses on each proficiency level in general pedagogical knowledge at second occasion (GPK2) and quality of courses in higher education scales (AMEs, with Bonferroni-Holm corrected p values)

	Level I. GPK2 Simple cognitive processes, practical language level	Level II. GPK2 Simple cognitive processes, professional/scientific language level	Level III. GPK2 Complex cognitive processes, practical/professional language level
OTL structured teaching (seminar)	0.025*	0.042**	− 0.003
OTL cognitive activation (seminar)	0.037**	0.040**	0.003

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$

Table 12 Findings from binary logistic regression analyses on each proficiency level in general pedagogical knowledge at second occasion (GPK2) and teaching practice scales (AMEs, with Bonferroni-Holm corrected *p* values)

	Level I. GPK2 Simple cognitive processes, practical language level	Level II. GPK2 Simple cognitive processes, professional/ scientific language level	Level III. GPK2 Complex cognitive processes, practical/ professional language level
OTL planning	0.006	0.02	0.016*
OTL teaching	0.025*	0.041*	0.018**
OTL linking	0.031	0.04*	0.003
OTL reflecting	− 0.007	0.039*	0.002

$^*p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$

example, if the "structured teaching" scale increases by one standard deviation, the probability of achieving the second proficiency level increases by 4.2%. Regarding the achievement of the highest level, the effect sizes of the scales are very small (0.3% and − 0.3%) and statistically not significant. From the data in Table 12, it is apparent that only one teaching practice OTL scale ("teaching") is significant for attaining the first level and three scales ("teaching," "reflecting," and "linking") for attaining the second proficiency level. Concerning the highest level, only the scales with direct instructional relations "planning" and "teaching" are statistically significant. Although their effect sizes are comparatively low at 1.6% and 1.8%, these two scales are the only quality and teaching practice OTL integrated in our analyses with an impact on attaining the most advanced level. The content scale exhibits significant effects in all models for attaining the first and second level (see Tables B1-B6 of the Electronic Supplementary Material) Thus, pedagogical content aspects play an essential role. However, whereas the scales "teaching" and "planning" affects the achievement of the highest level, the content scale does not.

While the scales "planning" and "reflecting" indicate no significant effects in ordinal regression analyses, in binary regression analyses, significant effects were found. This seems to be a contradiction only at first sight. Due to the stringent assumptions of the POM, differentiated analysis for each level may produce slightly different results. While the binary models perform different slopes for each level, the POM provides only one slope across all levels. Smaller effects might be therefore undetected. Accordingly, Bender and Grouven (1998) conclude that "for model checking and model building separate binary regression models are required in any case" (Bender and Grouven 1998, p. 814). This underlines the need to estimate not only POM but also separate binary models in order to obtain more detailed results.

4 Conclusion and discussion

Our analyses aimed to examine the differentiated effects of program characteristics (OTL) on the achievement of program outcomes (GPK). Currently, research in both fields OTL of pre-service teachers and proficiency levels in GPK is relatively rare. Therefore, the authors set a proficiency level model into focus and, additionally, put particular emphasis on the measurement of OTL. Furthermore, as longitudinal data are scarce, the authors used a sample of 1451 bachelor student teachers from Germany and Austria surveyed at two time points to enable examinations of changes over time.

Regarding our first research question, the results indicate that OTL contribute to the acquisition of competences and thus confirm previous findings (König and Klemenz 2015, König et al. 2017a, 2017b). However, one unanticipated finding was that only four of eight OTL scales show significant impacts across all levels (POM).

Concerning our second question, the impact of instructional quality was examined in more detail. We aimed to find out whether different learning opportunities led to differential effects on the achievement of certain proficiency levels. As the authors had hypothesized, both facets of instructional quality in seminars "structured teaching" and "cognitive activation" have an effect on the first two levels only, but not on the highest level. In line with theoretical assumptions (see "Section 1.2"), courses in higher education appear to contribute particularly to a theoretical-formal knowledge represented by the lower proficiency levels characterized by theoretical concepts and terms. The highest proficiency level, however, consists of tasks demanding complex cognitive processes. It requires taking multiple perspectives to solve problems and creating different options for action (Klemenz and König 2019). One can cautiously assume that higher education courses imparting theoretical knowledge do not necessarily appear to be the most appropriate OTL for achieving this level. Rather, it seems that in addition to courses in the academic setting also learning opportunities aiming at teaching practice and reflection are relevant to develop practical knowledge as part of professional knowledge (e.g., Berliner 2004; Schön 1983). This is also underlined, theoretically, by the example of the acquisition of classroom management expertise (see "Section 1.2") and, empirically, by our findings regarding our third hypothesis.

The third question focused on the assumption that in-school OTL in particular have an impact on achieving the highest proficiency level. The findings meet our assumptions partially. They indicate, interestingly enough, that the instruction-related OTL "planning" and "teaching" are the only of the teaching practice and quality facets assessed in our analysis with significant impacts on the most advanced level. However, their effect sizes are relatively low. The findings regarding the highest proficiency levels therefore need to be interpreted with caution. Nonetheless, given that practical knowledge is considered as action-oriented knowledge (Berliner 2004), our findings can be interpreted as follows: to achieve a proficiency level characterized by complex cognitive processes that demand at least proportionately practical knowledge, gaining experience directly related to classroom situations appears to be necessary. This is underlined by content-related OTL which play an important role in achieving lower levels, but not for the highest level. Our interpretation is in line with findings that emphasize the role of teaching practice during teacher education for the development of professional knowledge (König and Blömeke 2012; Blömeke et al. 2012; Schmidt et al. 2011b). Helsper (2001) assumes that practical knowledge cannot be achieved through theoretical knowledge or theoretical reflection, but only through an introduction to the teacher's actions themselves, through experience in practice and thus the acquisition of a practical habitus. However, one should also point out that more teaching practice does not automatically support the progress of professionalization. König and Rothland

(2018) highlight that practice should not be taken over without reflection. The reflection of teaching practice on the basis of theoretical knowledge is absolutely fundamental to ensure the teachers' behavior on a scientific basis. Therefore, the prerequisite for well-founded practical action in the teaching profession is, beyond subjective experience, first an understanding of the practice, its conditions, and the reasons why aspects can be appropriate or not in professional situations (Rothland 2016). This is reflected in the proficiency model.

Furthermore, even if the validation of the proficiency model was not the main focus, the results can be considered positive regarding criterion validity: as theoretically assumed, the participation in teaching practice and the positive rating of higher education courses correlate positively with the achievement of higher levels.

5 Limitations and implications

Although the findings seem to be promising, limitations have to be discussed as well. The item attributes have been carefully derived from theoretical approaches. However, in order to achieve a more differentiated description of the levels, it may be useful to develop further features and to relate them to further items. In addition, although the regression model applied explains a large part of the variance, there is still a large part unexplained. Not least also due to a lack of specific research, the applied method had to be referred to ex post and not a priori (Hartig 2007), i.e., the item attributes were generated after test construction. An a priori approach, which could be applied in future research building on the present study, is assumed to explain more variance (Sinharay et al. 2011) and is regarded to have more power to validate the interpretation of test scores (Hartig 2007). Moreover, further steps to validate the proficiency model, e.g., by examining relationships with additional external criteria, would be appropriate.

In addition, although in the present study, much effort has been invested in the development of items and scales to enable measurement in a low-inference way, this is still a field in which further progress is needed. In particular, the instructional quality scales could be revised to allow a more detailed assessment or analysis. Moreover, the authors assumed that the student teachers would be able to recollect accurately the quality facets of lectures and seminars as well as the different actions of teaching practice they carried out. However, since we used self-reports, it is possible that the students' statements are biased. Since this limitation applies to many studies, further efforts should be made to use other sources of data on OTL (e.g., observational data). Furthermore, only a small number of students ($n = 44$) achieved the highest level of competence in the fifth bachelor semester, which might be related to the limited scope of our analysis. We assume that advanced students in their master degree will attain the highest proficiency level more often. Findings from another study (König et al. 2018a) which applied the present proficiency level model on a sample of master students from three German universities emphasize this with about 35% of the student teachers reaching the most advanced level. Subsequent analyses could also draw even more attention to the differences on the program level. Overall, these limits should be an impetus for further research in the field of criterion-referenced testing and OTL.

However, despite the limitations, it should be mentioned that the authors are not aware of any study that has carried out such differentiated analyses at different knowledge levels. Thus, this project is the first comprehensive investigation of student teachers' OTL on certain GPK levels, which, beyond that, used longitudinal data. Therefore, further research should be undertaken to investigate the effects of program characteristics on detailed and direct measured outcomes in higher education such as proficiency levels in GPK.

What implications can be derived from the results for student teacher preparation programs? We may summarize that both the acquisition of theoretical-formal knowledge, which is apparently acquired particularly through courses at higher education institutes and practical knowledge, which is apparently acquired particularly in teaching practice, are fundamental components in the preparation of highly qualified teachers. However, teacher education institutions should be aware of the different effects of OTL on the achievement of professional knowledge and involve the findings in further decisions for curriculum planning. Particularly, against the background of differences in qualities of knowledge between novices and experts (Berliner 2004), precise adaptations of OTL to the respective level of expertise seem to be advisable. It obviously matters which OTL pre-service teachers make use of and during what phase of training or level of development of knowledge they do so. In addition, it could be shown that proficiency level models allow a detailed description of competence levels and the assessment which persons attain certain levels. These features make them attractive for licensing procedures and the evaluation of educational standards.

References

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). New York: Wiley.

Altrichter, H., & Posch, P. (2007). *Lehrerinnen und Lehrer erforschen ihren Unterricht: Unterrichtsentwicklung und Unterrichtsevaluation durch Aktionsforschung*. Bad Heilbrunn: Klinkardt.

Anderson, J. R. (1982). Acquisition of cognitive skills. *Psychological Review, 89*(4), 369–406.

Arnold, K.-H. (2014). Unterrichtsversuche als allgemeindidaktische Lerngelegenheit: Eine vergleichende Curriculumanalyse. In K.-H. Arnold, A. Gröschner, & T. Hascher (Eds.), *Schulpraktika in der Lehrerbildung. Theoretische Grundlagen, Konzeption, Prozesse und Effekte* (pp. 63–86). Waxmann: Münster.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*, 469–520.

Baumert, J., & Kunter, M. (2013). The COACTIV model of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers. Results from the COACTIV project* (pp. 25–48). New York, NY: Springer.

Beaton, E., & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*, 191–204.

Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology, 51*(10), 809–816.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*(5), 463–482.

Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society, 24*, 200–212.

Best, H., & Wolf, C. (2012). Modellvergleich und Ergebnisinterpretation in Logit- und Probit-Regressionen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie, 64*, 377–395.

Bieri, J., Atkins, A. L., Briar, S., Leaman, R. L., Miller, H., & Tripodi, T. (1966). *Clinical and social judgment*. New York: Wiley.

Blömeke, S. (2011). Forschung zur Lehrerbildung im internationalen Vergleich. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (pp. 345–361). Münster: Waxmann.

Blömeke, S., & Kaiser, G. (2012). Homogeneity or heterogeneity? Profiles of opportunities to learn in primary teacher education and their relationship to cultural context and outcomes. *ZDM – The International Journal on Mathematics Education, 44*(3), 249–264.

Blömeke, S., Kaiser, G., & Lehmann, R. (2010). *TEDS-M 2008 - Professionelle Kompetenz und Lerngelegenheiten angehender Sekundarstufenehrkräfte im internationalen Vergleich*. Münster: Waxmann.

Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: what matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education, 28*(1), 44–55.

Blömeke, S., Hsieh, F. J., Kaiser, G., & Schmidt, W. H. (Eds.). (2014). *International perspectives on teacher knowledge, beliefs, and opportunities to learn. Advances in mathematics education*. Dordrecht: Springer Science + Business Media.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: fundamental measurement in the human sciences (2nd)*. Mahwah, NJ: LEA.

Bromme, R. (1992). *Der Lehrer als Experte: zur Psychologie des professionellen Wissens*. Bern: Huber.

Bromme, R. (2001). Teacher expertise. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 15459–15465). Amsterdam, Netherlands: Elsevier.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Christensen, R. H. B. (2015). *Analysis of ordinal data with cumulative link models - estimation with the R-package ordinal.* https://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf. Accessed February 2018.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. Thousand Oaks, CA: SAGE.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). A NCME instructional module on setting performance standards: contemporary methods. *Educational Measurement, Issues and Practice, 23*, 31–50.

Clift, R. T., & Brady, P. (2005). Research on methods courses and field experiences. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education. The report of the AERA panel on research and teacher education* (pp. 309–424). Mahwah, NJ: Lawrence Erlbaum.

Cochran-Smith, M., Cannady, M., McEachern, K., Mitchell, K., Piazza, P., Power, C., & Ryan, A. (2012). Teachers' education outcomes: Mapping the research terrain. Teachers College Record, *114*(10), 1–49.

Cochran-Smith, M., & Villegas, A. M. (2016). Research on teacher preparation: charting the landscape of a sprawling field. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 439–547). Washington, DC: AERA.

Darling-Hammond, L. (2006). Assessing teacher education. The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*(2), 120–138.

DeMars, C. E., Sundre, D. L., & Wise, S. L. (2002). Standard setting: a systematic approach to interpreting student learning. *Journal of General Education, 51*, 1–20.

van den Akker, J. (2003). Curriculum perspectives: an introduction. In J. van den Akker, W. Kuiper, & U. Hameyer (Eds.), *Curriculum landscapes and trends* (pp. 1–10). Dordrecht: Kluwer Academic Publishers.

Embretson, S. E. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

ETS (2018). *Understanding teaching quality.* https://www.ets.org/research/topics/teaching. Accessed May 2018.

von Eye, A. (1999). Kognitive Komplexität. *Messung und Validität. Zeitschrift für Differentielle und Diagnostische Psychologie, 2*, 81–96.

Fenstermacher, G. D. (1994). The knower and the known: the nature of knowledge in research on teaching. *Review of Research in Education, 20*, 3–56.

Floden, R. (2015). Learning what research says about teacher preparation. In M. J. Feuer, A. I. Berman, & R. C. Atkinson (Eds.), *Past as prologue: the National Academy of Education at 50. Members Reflect* (pp. 279–284). Washington, DC: National Academy of Education.

Flores, M. A. (2016). Teacher education curriculum. In J. Loughran & M. L. Hamilton (Eds.), *International handbook of teacher education* (pp. 187–230). Dordrecht, the Netherlands: Springer.

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*(1), 1–56.

Good, T. L., & Brophy, J. E. (2007). *Looking in classrooms* (10th ed.). Boston: Pearson Education.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394–411.

Guerriero, G. (Ed.). (2017). *Pedagogical knowledge and the changing nature of the teaching profession*. Paris, France: OECD.

Harsch, C., & Hartig, J. (2011). Modellbasierte Definition von fremdsprachlichen Kompetenzniveaus am Beispiel der Bildungsstandards Englisch. *Zeitschrift für Interkulturelle Fremdsprachenforschung, 16*(2), 6–17.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 83–99). Beltz: Weinheim.

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63*, 43–49.

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

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72*, 665–686.

Hascher, T. (2014). Forschung zur Wirksamkeit der Lehrbildung. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (pp. 542–571). Münster: Waxmann.

Hattie, J. (2009). *Visible learning; a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Helmke, A. (2003). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze: Kallmeyer.

Helsper, W. (2001). Praxis und Reflexion. Die Notwendigkeit einer "doppelten Professionalisierung" des Lehrers. *Journal für lehrerInnenbildung, 1*(3), 7–15.

Houang, R., & Schmidt, W., (2008). *TIMSS international curriculum analysis and measuring educational opportunities*. 3rd IEA international research conference, 18–20 Sept 2008, Taipei, Chinese Taipei. http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2008/Papers/IRC2008_Houang_Schmidt.pdf. Accessed April 2018.

Jenßen, L., Dunekacke, S. & Blömeke, S. (2015). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. In Blömeke, S. & Zlatkin-Troitschanskaia, O. (Eds.), Kompetenzen von Studierenden. *Zeitschrift für Pädagogik*, 61. Beiheft (pp. 11-31). Weinheim U.A.: Beltz.

Jones, V. (2006). How do teachers learn to be effective classroom managers? In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice, and contemporary issues* (pp. 887–907). Mahwah: Lawrence Erlbaum Associates.

Kansanen, P. (2014). Teaching as a master's level profession in Finland: theoretical reflections and practical solutions. In O. McNamara, J. Murray, & M. Jones (Eds.), *Workplace learning in teacher education: international practice and policy* (pp. 279–292). Dordrecht: Springer.

Kauertz, A., Fischer, H. E., Mayer, J., Sumfleth, E., & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 135–154.

Klafki, W. (1985). *Neue Studien zur Bildungstheorie und Didaktik. Beiträge zur kritischkonstruktiven Didaktik*. Beltz: Weinheim.

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

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung im internationalen Vergleich. In E. Klieme & J. Baumert (Eds.), *TIMSS – Impulse für Schule und Unterricht* (pp. 43–57). Bonn: BMBF.

KMK (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. Accessed May 2018.

Kolbe, F.-U., & Combe, A. (2004). Lehrerbildung. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (pp. 853–877). Wiesbaden: VS.

König, J. (2009). Zur Bildung von Kompetenzniveaus im Pädagogischen Wissen von Lehramtsstudierenden: Terminologie und Komplexität kognitiver Bearbeitungsprozesse als Anforderungsmerkmale von Testaufgaben? *Lehrerbildung auf dem Prüfstand, 2*(2), 244–262.

König, J. & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften: Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerausbildung. *Zeitschrift für Erziehungswissenschaft, 12*(3), 499–527.

König, J. (2010). Lehrerprofessionalität - Konzepte und Ergebnisse der internationalen und deutschen Forschung am Beispiel fachübergreifender, pädagogischer Kompetenzen. In J. König & B. Hofmann (Eds.), *Professionalität von Lehrkräften - Was sollen Lehrkräfte im Lese- und Schreibunterricht wissen und können?* (pp. 40–105). Berlin: DGLS.

König, J., Blömeke, S., Paine, L., Schmidt, B. & Hsieh, F-J. (2011). General Pedagogical Knowledge of Future Middle School Teachers. On the Complex Ecology of Teacher Education in the United States, Germany, and Taiwan. *Journal of Teacher Education, 62*(2), 188–201.

König, J., & Blömeke, S. (2012). Future Teachers' General Pedagogical Knowledge from Comparative Perspective. Does School Experience Matter? ZDM - *The International Journal on Mathematics Education, 44*, 341–354.

König, J., & Seifert A. (Eds.). (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung*. Münster: Waxmann.

König, J. (2013). First comes the theory, then the practice? On the acquisition of general pedagogical knowledge during initial teacher education. *International Journal of Science and Mathematics Education, 11*(4), 999–1028.

König, J. (2014). Designing an International Instrument to Assess Teachers' General Pedagogical Knowledge (GPK): Review of Studies, Considerations, and Recommendations. Technical paper prepared for the OECD Innovative Teaching for Effective Learning (ITEL) - Phase II Project: A Survey to Profile the Pedagogical Knowledge in the Teaching Profession (ITEL Teacher Knowledge Survey). Paris: OECD.

König, J., Tachtsoglou, S., Darge, K. & Lünnemann, M. (2014). Zur Nutzung von Praxis: Modellierung und Validierung lernprozessbezogener Tätigkeiten von angehenden Lehrkräften im Rahmen ihrer schulpraktischen Ausbildung. *Zeitschrift für Bildungsforschung, 4*(1), 3–22.

König, J. & Klemenz, S. (2015). Der Erwerb von pädagogischem Wissen bei angehenden Lehrkräften in unterschiedlichen Ausbildungskontexten: Zur Wirksamkeit der Lehrerausbildung in Deutschland und Österreich. *Zeitschrift für Erziehungswissenschaft, 18*(2), 247–277.

König, J., Ligtvoet, R., Klemenz, S., & Rothland, M. (2017a). Effects of Opportunities to Learn in Teacher Preparation on Future Teachers' General Pedagogical Knowledge: Analyzing Program Characteristics and Outcomes. *Studies in Educational Evaluation, 53*, 122–133.

König, J., Bremerich-Vos, A., Buchholtz, C., Lammerding, S., Strauß, S., Fladung, I. & Schleiffer, C. (2017b). Modelling and validating the learning opportunities of preservice language teachers: On the key components of the curriculum for teacher education. *European Journal of Teacher Education, 40*(3), 394–412.

König, J., Darge, K., Klemenz, S. & Seifert, A. (2018a). Pädagogisches Wissen von Lehramtsstudierenden im Praxissemester: Ziel schulpraktischen Lernens? In J. König, M. Rothland & N. Schaper (Eds.), *Learning to Practice, Learning to Reflect? Ergebnisse aus der Längsschnittstudie LtP zur Nutzung und Wirkung des Praxissemesters in der Lehrerbildung* (pp. 287–323). Wiesbaden: Springer VS.

König, J., Doll, J., Buchholtz, N., Förster, S., Kaspar, K., Rühl, A.-M., Strauß, S., Bremerich-Vos, A., Fladung, I., & Kaiser, G. (2018b). Pädagogisches Wissen versus fachdidaktisches Wissen? Struktur des professionellen Wissens bei angehenden Deutsch-, Englisch- und Mathematiklehrkräften im Studium. *Zeitschrift für Erziehungswissenschaft, 21*(3), 1–38.

König, J. & Rothland, M. (2018). Das Praxissemester in der Lehrerbildung: Stand der Forschung und zentrale Ergebnisse des Projekts Learning to Practice. In J. König, M. Rothland & N. Schaper (Eds.), Learning to Practice, Learning to Reflect? Ergebnisse aus der Längsschnittstudie LtP zur Nutzung und Wirkung des Praxissemesters in der Lehrerbildung (pp. 1–62). Wiesbaden: Springer VS.

Klemenz, S. & König, J. (2019). Modellierung von Kompetenzniveaus im pädagogischen Wissen bei angehenden Lehrkräften: Zur kriterialen Beschreibung von Lernergebnissen der fächerübergreifenden Lehramtsausbildung. *Zeitschrift für Pädagogik*, *65*(3).

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV.* Münster: Waxmann.

Lampert, M., & Ball, D. L. (1999). Aligning teacher education with contemporary K-12 reforms. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: handbook of policy and practice* (pp. 33–53). San Francisco, CA: Jossey-Bass.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction, 19*, 527–537.

Manly, C. A., & Wells, R. E. (2015). Reporting the use of multiple imputation for missing data in higher education research. *Research in Higher Education, 56*(4), 397–409.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305–322.

McKenney, S., Nieveen, N., & van den Akker, J. (2006). Design research from a curriculum perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 62–90). London: Routledge.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Muijs, D., & Reynolds, D. (2011). *Effective teaching: evidence and practice* (3rd ed.). London: Sage.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35*(2–3), 95–101.

Peterson, C., & Scott, W. A. (1983). Toward fundamental measurement of dimensionality. *British Journal of Social Psychology, 22*, 197–202.

Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C., & Reusser, K. (2007): Structure as a quality feature in mathematics instruction: cognitive and motivational effects of a structured organisation of the learning environment vs. a structured presentation of learning Contant. In: Prenzel, M. (Eds.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (pp. 102-121). Münster: Waxmann.

Rost, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik, 50*(5), 662–678.

Rothland, M. (2016). Der Lehrerberuf als Gegenstand der Lehrerbildung. In M. Rothland (Ed.), *Beruf Lehrer/ Lehrerin. Ein Studienbuch* (pp. 7-15). Münster u.a.: Waxmann/UTB.

Schaper, N. (2009). Aufgabenfelder und Perspektiven bei der Kompetenzmodellierung und -messung in der Lehrerbildung. *Lehrerbildung auf dem Prüfstand, 2*(1), 166–199.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Schmidt, W. H., Houang, R. T., Cogan, L., Blömeke, S., Tatto, M., Hsieh, F. J., & Paine, L. (2008). Opportunity to learn in the preparation of mathematics teachers: its structure and how it varies across six countries. *ZDM - International Journal on Mathematics Education, 40*(5), 735–747.

Schmidt, W. H., Blömeke, S., & Tatto, M. T. (2011a). *Teacher education matters. A study of the mathematics teacher preparation from six countries*. New York: Teacher College Press.

Schmidt, W. H., Cogan, L., & Houang, R. (2011b). The role of opportunity to learn in teacher preparation: an international context. *Journal of Teacher Education, 62*(2), 138–153.

Schön, D. A. (1983). *The reflective practitioner – how professionals think in action*. London: Temple Smith.

Scott, W. A. (1962). Cognitive complexity and cognitive flexibility. *Sociometry, 25*, 405–414.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*, 454–499.

Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training, 1*, 43–65.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333–354.

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Sinharay, S., Haberman, S. J., & Lee, Y. (2011). When does scale anchoring work? A case study. *Journal of Educational Measurement, 48*, 61–80.

Slavin, R. E. (1994). Quality, appropriateness, incentive and time: a model of instructional effectiveness. *International Journal of Educational Research, 21*, 141–157.

Sonmark, K., Révai, N., Gottschalk, F., Deligiannidi, K, & Burns, T. (2017). *Understanding teachers' pedagogical knowledge: report on an international pilot study.* OECD Education Working Papers, No. 159. Paris: OECD Publishing.

Tachtsoglou, S. & König, J. (2017). Der Einfluss universitärer Lerngelegenheiten auf das pädagogische Wissen von Lehramtsstudierenden. Zeitschrift für Bildungsforschung, *7*(3), 291–310.

Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher Education and Development Study in Mathematics (TEDS-M): policy, practice, and readiness to teach primary and secondary mathematics. Conceptual framework.* East Lansing, MI: Teacher Education and Development International Study Center, College of Education, Michigan State University.

Tatto, M. T., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., & Reckase, M. (2012). Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries. Findings from the IEA teacher education and development study in mathematics (TEDS-M). Amsterdam: IEA.

Terhart, E. (1991). Pädagogisches Wissen. Überlegungen zu seiner Vielfalt, Funktion und sprachlichen Form am Beispiel des Lehrerwissens. In J. Oelkers & H.-E. Tenorth (Eds.), *Pädagogisches Wissen. Beiheft der Zeitschrift für Pädagogik* (pp. 129–141). Beltz: Weinheim.

Tiffin-Richards, S. P., Pant, H. A., & Köller, O. (2013). Setting standards for English foreign language assessment: methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice, 32*(2), 15–25.

Vanderlinde, R., van Braak, J., & Hermans, R. (2009). Educational technology on a turning point: curriculum implementation in Flanders and challenges for schools. *Educational Technology Research & Development, 57*(4), 573–584.

Von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*, 115–124.

Voss, T., Kunina-Habenicht, O., Hoehne, V., & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft, 18*(2), 187–223.

Watt, H. M. G., & Richardson, P. W. (2007). Motivational factors influencing teaching as a career choice: development and validation of the FIT- choice scale. *Journal of Experimental Education, 75*, 167–202.

Weinert, F. E. (2001). Concept of competence: a conceptual clarification. In D. S. Rychenj & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–66). Seattle: Hogrefe & Huber.

Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung*. Bielefeld: wbv.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: multi-aspect test software [computer program]*. Camberwell: Australian Council for Educational Research.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Stefan Klemenz[1] **· Johannes König**[2] **· Niclas Schaper**[3]

[1] Faculty of Human Sciences, Department of Education and Social Sciences, Empirical School Research, Quantitative Methods, University of Cologne, Gronewaldstraße 2a, Gebäude 214, Raum 1.18 (818), 50931 Köln, Germany

[2] Faculty of Human Sciences, Department of Education and Social Sciences, Empirical School Research, Quantitative Methods, University of Cologne, Gronewaldstraße 2a, Gebäude 214, Raum 1.16 (816), 50931 Köln, Germany

[3] Faculty of Arts and Humanities, Work and Organizational Psychology, Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany