

Structuring lessons as an aspect of pre-service teachers' planning competence: A scaling-up analysis

Matthias Krepf  · Johannes König 

Received: 15 September 2021 / Revised: 18 July 2022 / Accepted: 22 August 2022 / Published online: 16 September 2022
© The Author(s) 2022

Abstract Lesson planning is an essential part of teachers' daily work. In this study, we focus on *structuring* as an aspect of lesson planning, which generally can be defined as a clear, recognizable organization of instruction into individual phases and segments in which the teacher gradually builds up the complexity of the knowledge to be acquired and ensures a smooth flow of instruction through appropriate sequencing. In a previous study (Krepf and König in press), we conceived structuring as an aspect of lesson planning. To test the validity and reliability of this study's findings, a scaling-up study was conducted to determine whether structuring as an aspect of planning could be modelled reliably using a different and larger sample. In this study, 310 written lesson plans created by pre-service teachers during induction (172 at T1 [first lesson plan]; 138 at T2 [last lesson plan/state examination]) from North Rhine-Westphalia (NRW) and Berlin derived from the *PlanvoLL-D* project (König et al. 2020a, 2020b) comprised the study's data. The lesson plans were evaluated through content analysis using deductively formed categories. Afterward, the coding was quantified and analyzed using item response theory (IRT) scaling. The results indicated that two subscales could be separated in terms of content: a "contextualization" scale and a "phasing" scale. Furthermore, three explication levels could be distinguished. Measures of lesson structure planning increased during induction significantly with practical relevance. This study contributes to the research on modelling and measuring pre-service teachers' planning competence.

Keywords Lesson planning · Lesson structure planning · Competence · Induction · Teacher education

Dr. Matthias Krepf (✉) · Prof. Dr. Johannes König
Empirical School Research, University of Cologne, Gronewaldstr. 2, 50931 Cologne, Germany
E-Mail: matthias.krepf@uni-koeln.de

Prof. Dr. Johannes König
E-Mail: johannes.koenig@uni-koeln.de

Strukturierung von Unterricht als Aspekt der Planungskompetenz von angehenden Lehrpersonen: Eine Scaling-up Analyse

Zusammenfassung Die Planung von Unterrichtsstunden ist ein wesentlicher Bestandteil der täglichen Arbeit von Lehrpersonen. In dieser Studie fokussieren wir auf die Strukturierung als einen Aspekt der Unterrichtsplanung. Allgemein kann unter Strukturierung als Planungsaspekt eine klare, erkennbare Gliederung des Unterrichts in einzelne Phasen und Abschnitte verstanden werden, bei der die Lehrperson schrittweise Komplexität der zu lernenden Inhalte aufbaut und durch eine angemessene Sequenzierung für einen reibungslosen Unterrichtsablauf sorgt. In einer vorangegangenen Studie (Krepf und König im Druck) wurde Strukturierung als ein Aspekt der Unterrichtsplanung modelliert und empirisch geprüft. Zur Prüfung der Gültigkeit und Stabilität der Befunde soll nun mittels einer *Scaling-up-Studie* geprüft werden, ob sich Strukturierung als Planungsaspekt auch anhand einer anderen, größeren Stichprobe modellieren lässt. In der vorliegenden Untersuchung bilden 310 *schriftliche Unterrichtsplanungen von Referendar*innen* (172 zu T1 [erste Lehrprobe]; 138 zu T2 [letzte Lehrprobe/Staatsprüfung]) aus Nordrhein-Westfalen (NRW) und Berlin aus *PlanvoLL-D* (König et al. 2020a, 2020b) die Datengrundlage. Die schriftlichen Unterrichtsplanungen wurden inhaltsanalytisch mithilfe deduktiv gebildeter Kategorien kodiert. Anschließend wurden die Kodierungen quantifiziert und mittels Item-Response-Theorie (IRT) Skalierung ausgewertet. Die Ergebnisse zeigen, dass zwei Subskalen inhaltlich getrennt werden konnten: eine Skala „Kontextualisierung“ und eine Skala „Phasierung“. In formaler Hinsicht konnten drei Explikationsstufen unterschieden werden. Darüber hinaus zeigten die Ergebnisse, dass es im Laufe des Vorbereitungsdienstes zu einem Zuwachs an gemessener Planungskompetenz unter dem Aspekt der Strukturierung kam. Die Studie leistet einen wichtigen Beitrag zur Modellierung und Messung der Planungskompetenz von angehenden Lehrpersonen.

Schlüsselwörter Unterrichtsplanung · Strukturierung · Kompetenz · Referendariat · Lehrer*innenausbildung

1 Introduction

Structuring has emerged as a key feature of effective instruction in several empirical studies and meta-analyses (e.g., Brophy 2000; Hattie 2009; Seidel and Shavelson 2007; Walberg and Paik 2000). Generally, structuring as a planning aspect is a clear, recognizable organization of instruction into separate phases and sections in which the teacher gradually builds the complexity of the knowledge to be acquired and ensures a smooth flow of instruction by sequencing it appropriately (Doyle 2006; Kounin 1970).

However, only few empirical studies have examined modelling and measured planning competence (e.g., Cochran-Smith and Villegas 2016), particularly structuring (König et al. 2015). Only recently have projects emerged that focus on modelling and assessing teachers' planning competence in a standardized manner. Representa-

tive projects include *PlanvoLL (Planning Competence of Teachers [Planungskompetenz von Lehrerinnen und Lehrern]*, König et al. 2015) and *PlanvoLL-D (The Role of Professional Knowledge of Pre-Service German Teachers in their Lesson Planning [Die Bedeutung des professionellen Wissens angehender Deutschlehrkräfte für ihre Planung von Unterricht]*, König et al. 2020a, b).

Following this work, Krepf and König (in press) investigated in their study whether structuring as a *situation-specific skill* of planning competence could be measured and modelled while maintaining validity and reliability. In their study, written lesson plans created by more than 100 pre-service teachers in Berlin were analyzed at two measurement points from the *PlanvoLL project* (König et al. 2015). The lesson plans were evaluated based on a newly developed category system. Under the current research methodology, the category system comprised two different structuring factors: (1) a *context factor* and (2) a *process factor*. Furthermore, the explication levels of the planning decisions made (*naming, reasoning, and linking*) also were examined. The coding results then were quantified and assessed using IRT scaling (Rasch scaling).

The results indicate reliability and validity of the test instrument. However, the question remains as to how robust these findings are and whether they can be replicated. Replication studies are necessary to safeguard observed findings against chance (Rost 2007). Findings can only be viewed as validated if several studies on the same research question arrive at similar results (Döring and Bortz 2016). Therefore, a scaling-up study to examine whether structuring as a dimension of planning can be modeled reliably by using a larger sample. In our replication study, the data base comprises 310 written lesson plans from pre-service teachers during induction in North Rhine-Westphalia (NRW) and Berlin. Based on these data, various analyses (reliability and IRT scaling) will be conducted, presented, and subsequently discussed.

2 Theoretical background

2.1 Planning as a dimension of teacher competence

Broad agreement exists that *knowledge* is a key component of teachers' professional competence (Baumert and Kunter 2006; Shulman 1986). In a number of empirical educational studies, researchers have distinguished between general pedagogical knowledge, content knowledge, and pedagogical content knowledge (e.g., Baumert and Kunter 2006). In contrast, research on lesson planning skills is scarce, and a shared (competence) concept of lesson planning cannot be identified (König et al. 2015). One reason might be that lesson planning depends on the teaching *context* (John 2006). For example, teachers usually plan their lessons for a specific learning group. How they plan a concrete lesson also depends on the classroom routines and organizational structures with which a specific learning group is familiar with.

Because situation-specific aspects of teaching are relevant, many researchers today agree on an extended understanding of teacher competence. In suggesting a model of "competence as a continuum" (Blömeke et al. 2015), competence can be

characterized as existing at two poles: latent dispositions (cognitive and affective-motivational skills) at one pole and manifest performance (observed behavior) at the other. Blömeke et al. (2015) distinguished among (underlying) disposition (teacher knowledge) and situation-specific skills, with the latter assumed to be more proximal to classroom teaching (performance) than teacher knowledge as discussed in previous competence models (Baumert and Kunter 2006).

In terms of this classification, we view planning competence as a *situation-specific skill* comprising perception, interpretation, and decision-making (detailed König et al. 2020a, b). Planning competence develops in a planning situation that is characterized, in turn, by various conditions.

2.2 Theoretical modelling of structuring as an aspect of planning competence in PlanvoLL-D

Lesson planning is a central task of teachers (Baumert and Kunter 2006); therefore, it is an important part of the first phase of university teacher education (Hohenstein et al. 2014) and during the second phase of teacher training (induction phase, *Referendariat, Vorbereitungsdienst*). As part of this training, pre-service teachers should receive scientific and practical training in lesson planning. One focus concerns imparting knowledge on lesson planning (Rey et al. 2018).

Extant research has demonstrated that pre-service teachers have difficulties in planning lessons (Calderhead 1996). One reason is that they lack the broad knowledge base necessary for planning, as well as the flexibility to react appropriately to the unexpected. While experienced teachers take on a holistic approach to teaching, pre-service teachers often view teaching as a chronological sequence of events connected only partially.

Westermann (1991) found that significant differences exist between experts and novices in lesson planning. While novices tend to be oriented toward specific instructional objectives in developing structured lessons, experts place great emphasis on adopting the student perspective and orienting their planning toward students' specific (learning) needs. This orientation toward students' needs in lesson planning is linked closely to the aspect of structuring lesson content. Considering that the knowledge to be acquired is usually too complex for the students, it must be processed in a structured manner regarding both content and form.

Structuring is both a central aspect of effective teaching (see, e.g., Brophy 2000; Muijs and Reynolds 2011; Walberg and Paik 2000) and a core issue of general didactics and empirical teaching-learning research (see König 2012). In line with central didactic models and findings from empirical classroom research, at least two meanings of *structuring* can be distinguished (e.g., Doenau 1987; Kleickmann 2012; Lipowsky 2015): (1) *content* structuring (unit contextualization) and (2) *procedural/organizational* structuring (phasing).

2.2.1 Content structuring (unit contextualization)

In the context of lesson planning, the specific lesson to be presented is not only planned, but also located in a larger overall context to demonstrate how individual

lessons are interrelated (John 2006; Yinger 1980). In order to help students building a *well-organized knowledge base*, it is necessary to assess their prior knowledge and to decompose the lesson unit's content (not just for one lesson) into appropriate sub-parts (Leinhardt and Greeno 1986; Muijs and Reynolds 2011). These sub-parts must be connected (e.g., Drollinger-Vetter 2011).

The subdivision of the learning process into sub-parts and the connection of these usually is done from a temporal perspective (Scholl and Plöger 2020; Shavelson 1983; Yinger 1980). A lesson's outcome should build on previous lessons to allow for cumulative knowledge growth because knowledge, skills, and understanding usually emerge gradually through interaction among several individual lessons, forming a *unit* (Shavelson 1987). Such units are spread over different periods, usually one to four weeks (Borich 2004). The unit contains the topics to be covered in each lesson (Plöger 2008). When planning a unit, teachers must determine the subject area and material relevant to learning and formulate learning objectives (Maier 2012). Teachers also must select the content and objectives of the unit according to individual students' needs (Jäger and Maier 2019; Shulman 1987). To do this, teachers must consider students' knowledge levels and integrate their needs into the planning process (Berliner 2004; Kagan 1992). The learning content and associated learning objectives are distributed over several lessons with a decisive impact on content and method of individual lessons. Thus, the individual lesson is "part of a larger system of interrelated learning" (Borich 2004, p. 126). Therefore, the requirement for "unit contextualization" reflects that students often are expected to accumulate knowledge in a gradual, structured, or cumulative manner (Leinhardt and Greeno 1986; Muijs and Reynolds 2011).

Therefore, the written lesson plan should clarify the *topic* chosen for the unit and how this *topic* will be organized into subtopics. Furthermore, the individual lesson must be embedded in this larger context (unit). The lesson plans should specify what is expected of students at a particular point within the unit or what they are expected to accomplish, that is, what content will be discussed and how the content aspects relate to each other. Both the unit and each of the lessons need a clear, coherent thematic structure aiming at facilitating student's learning (Seidel et al. 2005). Three essential aspects should be elaborated upon:

1. In accordance with the German didactic tradition (for an overview, see Arnold and Koch-Priewe 2011), a content analysis should be conducted first to select a topic and legitimize it with reference to the curriculum (Jäger and Maier 2019). The teacher then should think about how this topic should be developed in terms of content within the framework of unit planning (regarding content).
2. The teacher must clarify each individual lesson's function in this unit, that is the individual lesson's position in the unit (Pauli and Reusser 2003). For example, an introductory lesson on a new topic will be designed differently from a lesson primarily intended to apply knowledge and transfer it to some other context. The teacher also should clarify how individual lessons will be connected, what function the lesson will have within the unit, how it will relate to preceding and subsequent lessons (Arnold and Koch-Priewe 2011), and whether the whole arrangement enables cumulative learning.

3. The content must be selected in a well-founded manner and prepared didactically; thus, a *didactic analysis* (Klafki 1995) is needed. This analysis serves as the didactic justification and legitimation of the (individual) lesson content. The teacher clarifies what content will be covered during the individual lesson and the new content's meaning for students. Added to justifying the content ("what"), the teacher must consider "how" the relevant facts and content can be taught to the students. Here, the "how" refers primarily to the naming, reasoning, and (at best) linking of methodological choices to one another. The various methodological decisions should be linked for a structural connection to emerge between content, methods, and lesson goals (see Heimann 1972; Klafki 2007; Maier 2012).

2.2.2 Procedural/organizational structuring (phasing)

Furthermore, teachers must consider how learning content will be developed during each lesson. In terms of the teaching process, the need for *procedural/organizational* structuring of the lesson includes dividing the lesson into phases, marking transitions, and naming students' behavior to be exhibited during transitions. Considering these aspects of the teaching process is necessary to maintain the lesson's flow (e.g., Doyle 2006; Kounin 1970).

For the flow of instruction to occur smoothly, the teacher must plan the organizational flow and individual student activities (Shavelson and Stern 1981; Yinger 1980). "Activity refers to the timing and sequencing (or pace or flow) of content and materials during instruction" (Shavelson 1987, p. 485). These activities and tasks should have "stimulating variety and optimal challenge that help students maintain their engagement in the task and minimize interruptions due to boredom or distraction" (Brophy 2000, p. 11). For momentum to be generated within an instructional activity, a teacher must anticipate before the lesson where momentum might be lost and think about how to prevent it. An essential aspect of this is the clear, recognizable structuring of the lesson into individual phases and sections and breaking down the lesson content into individual components. This includes the teachers' management of transitions between phases.

Lesson structure has become an important facet of instructional quality (Brophy 2000; Kyriakides et al. 2018). It is crucial to have a clear, identifiable structuring of instruction into discrete phases and sections in which the teacher ensures a coherent flow of instruction with minimal interruptions through proper sequencing (Muijs et al. 2014). Phasing typically refers to activity patterns (e.g., seatwork, presentation) that support the organization of instruction (e.g., Doyle 2006).

In written lesson plans, the phasing of the lesson is addressed in at least two places: in the lesson plan itself and in an articulation scheme. In the lesson plan, the didactic-methodological decisions can be described in detail. The articulation scheme (also progress plan) then reflects the condensed decisions, usually in tabular form. It puts teaching-learning processes into a chronological sequence. Articulation scheme plans not only have an orientation function, but also present an overview of "the temporal, factual, and social order of the planned lessons" (Plöger 2008, p. 170).

2.3 Previous study

A previous study (Krepf and König [in press](#)) used the data from the *PlanvoLL* project and examined if structuring could be modelled as an aspect of lesson planning. In view of the short theoretical explanations of the different meanings of structuring (see Sect. 2.2), we designed a category system that comprised the two factors *contextualization* and *phasing*.

Additionally, the category system is *hierarchically* structured in formal terms because lesson planning is a systematic, extensive *decision-making process* (Zahorik [1970](#)). The (formal) hierarchy includes the explication level of the planning decisions taken. At the lowest level (naming), it is checked whether relevant information is included in their plans. The middle level (reasoning) entails checking whether the decisions taken are reasonable. At the highest level (linking), it is a question of whether and how the individual decisions made are related to each other.

Table 1 presents the eight categories (items) for the scale of *contextualization*. The categories describe the content that links the individual lesson with the unit.

Since the explication levels of the planning decisions take on the lowest level (naming), the topics of the unit and the lesson are only named and derived from the framing standards.

At the second level (reasoning), the structure is described and explained. Indicators for reasoning could be the explanations of what happened within the unit, i.e., which content already had been discussed and how the content aspects are related to each other. At least two statements must be made: first, how the unit is structured, and second, what the function of the lesson to be planned is in this unit.

At the highest level (linking), didactic decisions are linked together so that a coherent overall picture can emerge. Therefore, a didactic analysis should be conducted (Klafki [1995](#)). The didactic analysis deals with the didactic interpretation, justification, and structuring of the content in relation to the specific lesson. It therefore should clarify what the content of the lesson is and what relevance the content has for the students. In addition, the methodological decisions need to be related to the content to clarify how appropriate facts and content are presented to the students.

The scale of *phasing* includes five upper categories and 15 sub-categories (items) to capture the phased structure of the lesson being planned and how these phases are connected (see Table 2).

As in our description of the contextualization facet of lesson planning, we distinguish between three explication levels. Indicative of the lowest level (naming) would be the mentioning of an articulation scheme (PN1). The following categories (of the naming level) focuses on the lesson's phase structure. The function of dividing the lesson into phases is to provide meaningful lesson phases and, thus, represent factual and temporal significance as a process structure (Plöger [2008](#)). Myriad phase models can be found in the literature and in empirical studies. These phase models provide a basic framework through which instruction can be structured. Although a three-phase structure often is used in larger international studies, such as TIMSS 1999 (Review—Introducing New Content—Practicing/Applying/Consolidation; Hiebert et al. [2003](#)) or in TEDS-M 2008 (König and Blömeke [2009a](#)), and the three-phase structure represents a *minimal consensus* that can be derived from U.S. and German-

Table 1 Contextualization (C) subscale categories (Krepf and König [in press](#))

Explication level	Category description	Code	Text sections from different lesson plans
Naming (N)	The lesson plan for the unit contains references. The lesson's topic and the unit's topic are <i>named</i> . Ideally, the topic is not only named, but also derived from the framing standards	CN1	<i>Unit topic</i> : "Fables: Animals and their action—how they teach people about life." <i>Lesson topic</i> : "Together we are strong! We'll solve the problem"—Developing the lesson of the fable "The Two Donkeys" through a change in perspective from the learner to the donkey in the form of dialogue, as well as a transfer to one's own life world
	In the lesson plan, a distinction is made based on the lesson's content focus. The lesson's focus is <i>named</i> (e.g., introduction lesson, in-depth lesson, exercise lesson). A paraphrase of the lesson focus is sufficient	CN2	The lesson focuses on consolidating and applying the knowledge of rendering literal speech through indirect speech
Reasoning (R)	To help students building a well-organized knowledge base, the unit's overall theme is broken down into appropriate sub-aspects. Which content has been learned so far in the unit and how the previous content and new content to be learned relate to each other are clarified	CR1	"The whole unit is designed to allow <i>problem-based learning</i> . During the present lesson's introductory phase, stimulated by a picture that shows Faust and Margarete meeting for the first time, students are asked to name the leading question introduced and worked on during the unit's previous lesson. In the lesson today, this leading question will be examined and worked on during the working-on-the-topic phase collaboratively, following the introductory phase. The reference to the picture stimulus that already was applied in the previous lesson ensures that a connection exists to the unit's previous lesson."
	In the written lesson plan, justify the function of the lesson to be planned with regard to the unit. The lesson's focus is described in detail and reasoned (e.g., if it is an introduction lesson or an application lesson)	CR2	"This lesson is designed as a <i>practice and application lesson</i> . [...] In this lesson, the students should learn some of the mechanisms that contribute to a certain effect in the poem and use them to interpret the content. To do this, they identify one creative device at a time. [...] This procedure covers the standard used to 'assess the possible effect of basic creative devices and describe them in technical language' (new RLP, Part C, p. 26). Usually, a poem is interpreted first. Only during the second step is the poem's effect examined, particularly through compositional devices. During this lesson, this procedure is streamlined by focusing directly on the means of composition and the effect they produce. Building on this, the poem will be rewritten partially during the following lesson in the sense of action orientation to enable a more holistic approach."

Table 1 (Continued)

Explication level	Category description	Code	Text sections from different lesson plans
Linking (L)	In the lesson plan, the lesson is related to the unit (linked) and describes the lesson theme's relevance for the complete unit. Three aspects are necessary for this: First, how the unit is structured is explained. The content will be divided into clearly manageable sub-aspects. Second, the function of the lesson to be planned in this unit is described. Third, the unit and lesson are linked to each other	CL1	<p>Unit topic: Linguistic diversity in the 21st century. Examination of personal and societal language change.</p> <p>Lesson topic: Educating multilingually—but how?—Investigating the principles of multilingual education.</p> <p>The lesson's significance in the unit's context: After the learners have gained an overview of and critically commented on the theories of first language acquisition, they are given the opportunity to reflect on second language acquisition and their own multilingualism. This aims to make learners aware of this phenomenon's significant influence on their own lives. [...] Simultaneously, it should be made clear to the students that multilingualism is no longer an "exotic" exception, but rather the norm in today's society. In this respect, it is important to convey to the students that they, too, can help shape their social environment's "multilingual landscape" and actively use the knowledge they have gained about the acquisition of other languages. A prominent example is the education of their own children. After dealing with the advantages and disadvantages of multilingualism from the perspective of EU policy and neuroscience in the previous lessons, they already can assess multilingual education's relevance in the development of modern society. In the lesson to be presented, the learners will examine the question of how multilingual education can succeed, taking into account the advantages and disadvantages, so that they can recognize how possible courses of action must be adapted to a concrete life situation, and that basic principles contribute to the success of such education as a whole. After dealing with multilingualism in the family, the students can transfer their knowledge to other areas of life, e.g., the school realm</p>

Table 1 (Continued)

Explication level	Category description	Code	Text sections from different lesson plans
	In the lesson plan, detailed planning of the lesson takes place. The lesson's content is explained (subject analysis), and the lesson's subject matter is justified. Justifying the subject clarifies, among other things, a) the subject's importance for the students (present, future) and b) what is <i>typical</i> about this subject	CL2	Excerpt: "The main didactic focus is on the production of poems, the related gathering of inspiration, and the expression of individual moods. <i>Exemplarity</i> : Love poems and the associated preoccupation with moods in relation to different motifs are exemplary for an important, contrasting sub-area of poetry. <i>Relevance to the present</i> : For students at this age, their first experiences with love and relationships play an important role in their everyday lives. In this context, young people have certain ideas and may already have had individual experiences that they can express and reflect on by dealing with different love-lyrical motifs. Thus, the topic of the lesson series is linked to the young people's life worlds. <i>Future reference</i> : During the students' lives, they will successively expand their wealth of experience regarding love, relationships, and friendship. Furthermore, they will keep asking themselves what love means to them. Therefore, it also is essential to focus on this personal and sensitive topic in school."
	The methodical decisions (methods, social forms, and media) are justified. The justification views the content as learned. The method decisions are adjusted to the content and linked to them	CL3	"The students work in homogeneous partner groups during the elaboration phase because students who are rather reserved in other mathematics lessons are very motivated in geometry lessons and achieve good results. Partner work was chosen as the social form for the development phase to practice argumentation and to prepare the students for subsequent evaluation in the plenary."
	The competencies to be achieved are formulated (including learning objectives). The formulated competency statements are derived with reference to the framework curriculum. The objectives are related to the content and method decisions	CL4	"In differentiated group work, with different roles and individual responsibility, the students open up a non-fiction text by making targeted use of reading strategies and formulate questions for a class quiz in which the groups compete against each other by specifically searching for and identifying individual pieces of information in the text."

language literature on lesson planning, we initially assumed a *four-phase* structure. Based on Herbart's four-phase model, this still is used today in a modified form (Maulana et al. 2012). The phases—including introduction, working on the topic, and checking results and providing feedback—can be found in relevant textbooks for (pre-service) teachers (see Meyer 2004). In our case, we chose a four-phase lesson structure comprising the *introduction* (PN2P1), *working on the topic* (PN3P2), *checking results and providing feedback* (PN4P3), and *application/transfer* (PN5P4). Here, the coding of the phase structure is done in two steps: (1) In the first, following *TEDS-M* (König and Blömeke 2009a, b), the phase names were coded low-inferentially using a comprehensive coding guide that included different phase names deductively derived from the German- and English-language literature (e.g., Good and Brophy 2007; Kaiser and Kaiser 1991; Klingberg 1982; Meyer 1999), among others. (2) In the second step, the individual coded phase labels were assigned to the superordinate four-phase structure.

The next level (reasoning) indicates whether the functions of the individual phases have been described or reasoned (Rosenshine and Stevens 1986). For example, for the lesson introduction, “activation of prior knowledge” could be mentioned as a function, followed by an explanation of how this activation of prior knowledge will be achieved.

At the highest level (linking), the individual phases should be related to each other. Due to this exposed position of managing transitions (Arlin 1979; Doyle 2006), concrete information about how each activity (or phase) is linked should be found during instruction. We focused on simple transitions. For example, the transition from the introductory phase (social form: plenary) to the elaboration phase (social form: group work) can be addressed in one lesson plan. The focus should be on change in social form (e.g., “What rules apply to group work? Are certain

Table 2 Phasing (*P*) subscale categories (Krepf and König in press)

Explication level	Category description	Code	Text sections from different lesson plans
Naming (N)	An articulation scheme is mentioned. The scheme can be a table. The lesson's logical sequence is clear	PN1	–
	The articulation scheme contains the lesson's course phases. At least a four-phase structure is evident: <i>introduction</i> (P1); <i>working on the topic</i> (P2); <i>checking results and providing feedback</i> (P3); and <i>application/transfer</i> (P4). The individual phases are coded separately	PN2P1	In the lesson plan: “Preview of the lesson” Coded as: “introduction”
		PN3P2	In the lesson plan: “Elaboration” Coded as: “working on the topic”
		PN4P3	In the lesson plan: “Securing results” Coded as: “checking results and providing feedback”
		PN5P4	In the lesson plan: “Transfer” Coded as: “application/transfer”

Table 2 (Continued)

Explication level	Category description	Code	Text sections from different lesson plans
Reasoning (R)	The phases are described following their didactic function: (1) <i>introduction</i> (P1); (2) <i>working on the topic</i> (P2); (3) <i>checking results and providing feedback</i> (P3); and <i>application/transfer</i> (P4)	PR1P1	(1) Introduction: The first encounter with the poem should take place through an audio text, which is a change from the teacher's voice for the students, and the expression of spontaneous impressions activates and motivates the students. Afterward, the poem should be listened to again, but the students should orient themselves on guiding questions, which serve as a basis for the development phase and something to be taken up again at the end of the lesson. function: Activation/motivation and introduction to the topic
		PR2P2	(2) Working on the topic: During the elaboration phase, the representation of the city and the forest by the lyrical "I" will be elaborated on intensively. For this purpose, the students will now receive the text, which they should use to prove the representation. A division of labor (division of the class) is appropriate because the content areas of the forest and world are distinguished clearly from each other in the poem. Partner work was chosen because interpretation of the poem is very demanding in some places, and the students can support each other. The help cards and consolidation cards are used for internal differentiation. The help cards refer specifically to the tasks during the development phase. The consolidation cards serve as preparation for the next lesson. function: Active engagement with content
		PR3P3	(3) Checking results: Presentation by a pair of students (per "topic area") with the help of slides was chosen because this is a familiar procedure for the students from the lessons. The rest of the students will be active, as they may need to add to or correct the results. Furthermore, a reference back to the introduction should be made, and the students should check whether their initial assumptions about the mood of the lyrical "I" have been confirmed through the text work. function: Compiling and discussing the results
		PR4P4	(4) Application/Transfer: The deepening phase (application phase) follows on from the previous use of detailed knowledge in that the students now have to put this into action-oriented form. The letters also will show the depth of the individual students' understanding of the text and their detailed knowledge. All students have the opportunity to deepen their understanding of the text through partner work. function: applying and deepening the content learnt

Table 2 (Continued)

Explication level	Category description	Code	Text sections from different lesson plans
Linking (L)	The corresponding transitions are described in the lesson plan. How the individual phases are linked with each other is made explicit. The individual transitions are first to second phase (P1–P2), second to third phase (P2–P3), and third to fourth phase (P3–P4)	PL1P1P2	“Before the working phase begins, the further course of the lesson is visualized on a flipchart, and the work assignments are clarified to create transparency. In the first assignment, the form of interpretive writing is used to make the students’ intuitive understanding of the perceptions of the lyrical speaker in the poem visible to prepare or initiate subsequent analysis. Working on this task is made easier by the introduction because the students can better put themselves in the place of the lyrical speaker of the poem through the preparatory work.”
		PL2P2P3	“The (work) on the topic phase is designed for collaboration among students: Three groups of students explore the same text section. These findings are brought together in the subsequent presentation phase. Considering that all students are dependent on other groups’ findings, their achievement motivation will be enhanced.”
		PL3P3P4	“The lesson should be concluded with the transfer phase. During this phase, possible options for action—which the groups worked out first, then discussed during the securing phase and wrote on the board—should be applied. For this purpose, the scenic presentation of the introduction will be taken up again.”
	The corresponding transitions are described in the articulation scheme. How the individual phases are linked with each other is made explicit. The individual transitions are first to second phase (P1–P2); second to third phase (P2–P3); and third to fourth phase (P3–P4)	PL4P1P2	–
		PL5P2P3	–
		PL6P3P4	–

rules that apply to group work repeated/discussed?") and/or the didactic approach (e.g., formulation of the work assignment; anticipation of possible comprehension problems). Linking individual decisions can occur at two different points: in the formulated lesson plan and/or in the articulation scheme. Both presentation areas were recorded separately to gain information about the exact location in the lesson plan where such links are addressed.

In the previous study by Krepf and König (in press), 106 Berlin pre-service teachers took part. With exception of one, each of them provided two lesson plans, resulting in a total sample of 210 lesson plans from the two measurement time points. The written lesson plans were evaluated content-analytically using deductively formed categories (see Tables 1 and 2). Dichotomous codes (1 or 0) were given to indicate whether each criterion was met or not. A maximum of 23 possible points could therefore be scored. The respective coding events subsequently were quantified. IRT scaling was used to estimate item and person parameters (Rasch 1960). One-, two-, and three-dimensional IRT scaling analyses were conducted to test dimensionality, and different models were evaluated to determine which model better reflected the data. The three central results were (Krepf and König in press):

1. The scales *contextualization* and *phasing* could be separated empirically (scale contextualization EAP=0.66; Theta=1.49; scale phasing EAP=0.75; Theta=1.79). The two-dimensional modelling showed a better fit than one-dimensional modelling (Chi-square test between difference two *deviance* values = 88.14; $df=2$; $p<0.001$).
2. The scales were organized hierarchically. The hierarchization corresponds to the explication levels of the planning decisions made. The three explication levels could be separated empirically (subscale naming EAP=0.61; Theta=0.87; subscale reasoning EAP=0.73; Theta=1.96; subscale linking EAP=0.68; Theta=1.18). The difference between the two *deviance* values (40.91) was significant for $df=5$ ($p<0.001$), implying a better fit of the three-dimensional model.
3. An increase was demonstrated in measured planning competence under the aspect of structuring during induction phase. The score increased significantly ($t_{score}(103)=-5.77$, $p<0.001$) from the beginning (T1: $M=10.01$, $SE=0.34$, $SD=3.46$) to the end (T2: $M=12.46$, $SE=0.29$, $SD=3.01$) of the induction. An increase could also be observed for each of the three explication levels (NamingT1: $M=5.12$, $SE=0.11$, $SD=1.1$; NamingT2: $M=5.45$, $SE=0.1$, $SD=0.97$; ReasoningT1: $M=2.45$, $SE=0.16$, $SD=1.6$; ReasoningT2: $M=3.23$, $SE=0.13$, $SD=1.31$; LinkingT1: $M=2.44$, $SE=0.17$, $SD=1.8$; LinkingT2: $M=3.77$, $SE=0.16$, $SD=1.68$). By the end of induction, pre-service teachers are better able to explain their thoughts and provide explanations for their decisions ($t_{Nam}(103)=-2.41$, $p=0.02$; $t_{Rea}(103)=-4.39$, $p<0.001$; $t_{Link}(103)=-6.19$, $p<0.001$). The results to date suggest that a reliable and valid test instrument has been developed.

2.4 Research questions

The present study was designed as a replication study to test the validity and reliability of the findings of the previous study using a new and larger sample. Considering that a replication study is presented, the three research questions (RQ) of the previous study were investigated again. However, a fourth RQ was added, taking into account that pre-service teachers from two German states participated in the replication study.

RQ 1 Can structuring as an aspect of planning competence be measured and mapped reliably on a continuum within the IRT scaling framework? Is it a one- or two-dimensional construct?

RQ 2 Can the different explication levels (naming, reasoning, and linking) be mapped through an item-level analysis?

RQ 3 Does structuring as a part of planning competence increase over time (during induction)? Which explication levels (naming, reasoning, and linking) change throughout the induction and, thus, are a prominent aspect of competence development during the induction of pre-service teachers?

RQ 4 Can the empirically found structure of structuring as an aspect of planning competence be confirmed for different subsamples (NRW and Berlin)?

3 Method

3.1 Sample

In the project *PlanvoLL-D*, pre-service teachers who entered the second phase of teacher education (induction) were surveyed in two German states, North Rhine-Westphalia (NRW) and Berlin, in spring 2016. The sample (NRW) included pre-service teachers who attended a teacher education program to qualify for lower secondary school (*Haupt-/Real-/Gesamtschule*) or for lower and upper secondary school (*Gymnasium/Gesamtschule*). In Berlin, the corresponding teacher education qualifies pre-service teachers for lower and upper secondary school (*Integrierte Sekundarschule/Gymnasium*). In both states, these teacher education programs are the only ones that qualify teachers of German for secondary schools.

The data consisted of 310 written lesson plans from pre-service teachers from two measurement points: T1 at the beginning (first lesson plan) and T2 at the end (last lesson plan) of their internship. The time between T1 and T2 was approximately 1.5 years. At T1, 172 pre-service teachers participated. They were approximately 29 years old ($M=28.9$, $SD=4.9$). 85% were female; 35% (60 pre-service teachers) came from Berlin and were qualifying themselves to teach at Secondary Levels I and II; 65% (112) qualifying to teach at Secondary Level I came from NRW. 42% (73) qualifying to teach at Secondary Levels I and II, and 23% (39) were

qualifying to teach at Secondary Level I. At T2 138 pre-service teachers submitted a lesson plan (response rate: 80%). Out of this, a panel sample of 116 could be formed from pre-service teachers who submitted their lesson plans at both time points.

3.2 Instruments

Altogether, the evaluation instrument comprised 23 categories (items, see Tables 1 and 2). Regarding content, the two subscales *contextualization* (eight items) and *phasing* (15 items) are distinguished. Furthermore, three different explication levels (naming, reasoning, and linking) were taken into consideration.

3.3 The written lesson plan as an authentic planning document

Various methodological approaches have been established to measure lesson-planning competence (see Rothland 2021), including self-assessment procedures, vignette tests, planning knowledge tests, and analyses of authentic planning documents (written lesson plans). We focused on *written lesson plans*, which are a product of lesson planning (Grzesik 1979) and long have been an integral part of practical teacher training (Besa et al. 2020).

Lesson planning requires a series of decisions. Writing a lesson plan represents a mental effort to put these decisions into written form. Written lesson plans “document the teacher’s planning considerations; they provide information about the choice and justification of the topic, about the learning goals to be achieved, about the methods, media, and social forms used” (Plöger 2008, p. 170). The teacher must critically reflect on their own approach and justify the decisions (Esslinger-Hinz et al. 2013). These justifications should be coherent to external people (e.g., subject teachers) because written lesson plans are part of the grading process. Therefore, it should be described explicitly what will happen during the lesson. The lesson plan does not reflect the entire planning process (e.g., weighing alternatives), but they do contain the teacher’s final planning considerations. Thus, the written lesson plan represents a first product of lesson planning. The category system (instrument) should be used to reconstruct pre-service teachers’ planning decisions in written lesson plans.

3.4 Measures and data analyses

Data analysis was undertaken in two steps. First, the lesson plans were analyzed using the *qualitative content analysis* method (Mayring 2014) using the deductively formed category system. Overall, 23 categories were used, each of which was defined, as well as made concrete, using a prototypical text passage derived from written lesson plans (see Tables 1 and 2). These descriptions and text examples together formed a *category system* (Mayring 2014). Conducting the content analysis required a correct and precise application of each category. Therefore, it was necessary to prepare the coders for the coding process. Intensive practice periods were used to train the coders in applying the category system. At first, text passages

were coded together, then separately. Afterward, errors, discrepancies, and deviations were discussed. Once they gained sufficient confidence in using the categories, the coding process began.

Two trained raters independently coded the written lesson plans using dichotomous items. If a corresponding criterion was covered in the lesson plans, the raters were assigned a code of 1. A code of 0 was assigned if the criterion was not found in the written plan. If the written plans did not contain sufficient information on the analysis criterion, a value of 9 was assigned (at T1 one lesson plan was coded 9 throughout, so 171 lesson plans could be validly analyzed).

The category system used was not only an important prerequisite for the coding process itself, but also for the evaluation of the coding's stability and reproducibility. To ensure *intercoder reliability*, the two coders double-coded approximately 15% of the lesson plans, and these results were compared with each other. A *Kappa* between 0.60 and 0.75 was established in advance as a sufficient level for (intercoder) reliability (e.g., Fleiss and Cohen 1973). The overall Kappa was good ($\kappa > 0.8$). For the subscale phasing, the Kappa was 0.86. For the subscale contextualization, the kappa was slightly lower (0.71). Nevertheless, acceptable-to-good Kappa values for intercoder reliability were elicited, indicating reliable and stable coding.

Second, the coding registered was processed quantitatively. For each coded text passage, one point was awarded—like a correct answer on a test. The categories used were viewed as scored—like test items. Accordingly, the introduced category system contained a total of 23 categories (items) with which the lesson plans were examined: eight items for the subscale *contextualization* and 15 for the subscale *phasing*. Both scales can be interpreted—analogueous to item scales—as subscales of a construct containing a set of defined items. In accordance with this, the category system formed a scale that included 23 variables (categories) with two scales (contextualization and phasing) and three scales for explication levels (naming, reasoning, and linking).

Due to the size of the sample, the virtual case approach was used for scaling (e.g., Rost 2004). This is also known as “concurrent calibration” (von Davier et al. 2006), in which longitudinal subjects (i.e., subjects in which two measurements are available) represent two cases in a scaling file (one “real” and one “virtual”), thereby increasing the estimate's power (Bond and Fox 2007). The first measurement (at the beginning of induction) included 171 pre-service teachers' lesson plans. The second measurement included 138 pre-service teachers' lesson plans. The result was a scaling of 309 cases. With this increased number of cases, IRT scaling was possible ($n < 150$) (e.g., Bond and Fox 2007).

The coding was conducted with *dichotomous* items, so a *1-PL Rasch model* (IRT) was used (Rasch 1960). The data analyses were conducted using *ConQuest* (Wu et al. 1997), which assigns a difficulty parameter to each item based on its solution rate and an ability parameter to each subject according to demonstrated performance using a *maximum likelihood* procedure.

A major advantage of IRT scaling is that it allows for mapping of person ability scores and item difficulties to a common scale. Another advantage is that *ConQuest* provides information about the scale reliability and global quality of the models examined. Empirical reliability was calculated using multidimensional *expected a posteriori estimation* (EAP; de Ayala et al. 1995) which allows an unbiased

description of population parameters (Adams et al. 1997). It can be interpreted similar to Cronbach's Alpha (Rost 2004). Item fit statistics provide detailed insights into the quality of items and how they work in the assumed measurement model (Wright and Linacre 1994).

Furthermore, the deviation index (*deviance*; Wu and Adams 2006) compares the investigated models' global fit and provides information on which models provide a better fit to the data (*degree of goodness of fit*). Lower deviance indicates a better fit. To find dimensionality evidence in the data, one-dimensional, two-dimensional (the *contextualization* and *phasing* subscales), and three-dimensional (the explication levels *naming*, *reasoning*, and *linking*) IRT scaling analyses were conducted. Moreover, the models' *goodness of fit* (chi-square difference test) was examined to determine which model fit the data better.

To check the differences between explication levels, we used a *single factor analysis of variance* (ANOVA), then applied a *t-test for dependent samples* and calculated the effect size for the differences between measurement points 1 and 2 to answer our third RQ.

IRT scaling allows data from two measurement time points to be related to each other so that the *weighted likelihood estimates* (WLEs; see Warm 1989) found in the virtual cases approach can be exported as person parameters. Furthermore, we examined whether separate scaling was possible for both measurement time points. Separate scaling for each measurement time point allows for a comparison of the item difficulty parameters obtained. For this purpose, item difficulty parameters were correlated with each other (T1 and T2). High correlations then can be interpreted as indicating the presence of measurement invariance (see Bond and Fox 2007). We viewed the prerequisite for longitudinal scaling as met if a high correlation (above 0.7) is present. Moreover, we conducted a *regression analysis* for the two measurement time points. This included planning competence as the dependent variable and teaching type as the independent variable.

4 Results

4.1 RQ 1

First, a one-dimensional scaling was conducted within the framework of the virtual cases approach, resulting in acceptable values of the EAP reliabilities for the total scale (EAP=0.81; Theta=1.71), that is, the items sufficiently scatter across the ability spectrum. Table S1 (see Electronic Supplementary Material, ESM) provides the item statistics. The modelling reveals that different solution frequencies exist for the one-dimensional model: Some items were used very often (the naming of the phases), while others were found very rarely (the linking of the individual phases with each other).

During the scaling procedure, a difficulty parameter was assigned to each item based on its solution rate using a *maximum likelihood method* (estimates). Items with an estimate of less than 0 are classified as "too easy" because many pre-service teachers have "solved" this item. The *weighted mean squares* (WMNSQ)

Table 3 Findings from the IRT scaling analysis on content dimensions

Model	Deviance	Estimated parameters	Chi-Square difference
1-dimensional	5941.03	25	207.29 ($df=3$)
2-dimensional	5733.74	28	$p < 0.001$

(from 0.8 to 1.2; see, e.g., Adams and Wu 2002), as well as *t-values* (> -1.96 and < 1.96 ; see Bond and Fox 2007) largely fell within the expected range. Negative *t-values* indicated little variation in response patterns, which could be interpreted as an indication that a “minimal consensus” exists on the requirements for lesson plans.

Significant correlations were found regarding the total score to be achieved (T1+T2) and the two subscales, indicating a homogeneous construct (Table S2 in the ESM). The correlations also were calculated regarding the two measurement time points. The intercorrelation at T1 was 0.32** which is significant at the $p < 0.01$ level. No significant intercorrelation was found at T2 (0.11).

We performed two-dimensional scaling in the virtual cases approach. The EAP reliabilities for the two subscales lied within an acceptable range (subscale *contextualization* EAP=0.82; Theta= 5.79; subscale *phasing* EAP=0.78; Theta= 1.15). Next, we compared the one-dimensional and two-dimensional models, with a chi-square test performed to determine the best model. Table 3 provides the chi-square test results, which were significant for $df=3$, implying that the two-dimensional model is more suitable.

4.2 RQ 2

RQ 2 tests whether the differences between explication levels can be replicated (naming, reasoning, and linking). For scaling, we used the approach of virtual cases (see above) and performed three-dimensional scaling. The reliabilities of the three explication levels fell within an appropriate range (naming subscale: EAP=0.8; Theta=0.99; reasoning subscale: EAP=0.75; Theta=2.01; linking subscale: EAP=0.81; Theta=2.65). Next, we compared the one-dimensional scaling model's deviance with that of the three-dimensional model that assumes the three explication levels. The difference between the two deviance values (34.7) was significant for $df=4$, suggesting that the three-dimensional model provided a better fit than the one-dimensional model (see Table 4). Thus, the model comparison favors the three-dimensional model.

Table 4 Findings from an IRT scaling analysis of explication

Model	Deviance	Estimated parameters	Chi-Square difference
1-dimensional	5941.03	25	34.7
3-dimensional	5906.33	29	($df=4$) $p < 0.001$

Fig. 1 Item threshold parameters (*circles*) and means (*rectangles*) from one-dimensional IRT scaling split into the explication levels naming (*left side*), reasoning (*middle*), and linking (*right side*)

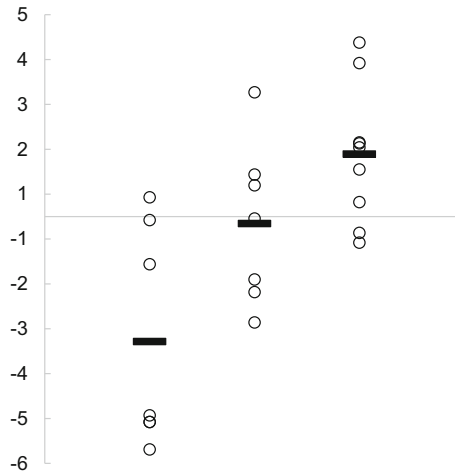


Fig. 1 provides the distribution of item threshold parameters from the unidimensional IRT scaling, with each item indicated by a circle. The distribution of items was divided according to the three explication levels. The rectangles in Fig. 1 are the specific mean values of the item parameter estimates for the three explication levels (Naming: $M = -2.78$, $SE = 0.92$, $SD = 2.43$; Reasoning: $M = -0.15$, $SE = 0.68$, $SD = 1.79$; Linking: $M = 1.39$, $SE = 0.51$, $SD = 1.53$). The results indicate that different explication levels can be differentiated in the data. Using a single-factor ANOVA, the overall mean difference between the three explication levels was found to be significant ($p < 0.001$). A post-hoc test (Bonferroni) indicated a highly significant difference ($p < 0.001$) between naming and linking, and a barely non-significant difference between naming and reasoning ($p = 0.055$). Therefore, we view our assumptions underlying RQ 2 as supported.

4.3 RQ 3

RQ 3 aimed to examine if structuring as a part of planning competence increased during the induction. Table 5 indicates that the solution frequencies for all items increased during the induction phase. Almost all the increases were significant (Wilcoxon test). Some items had very high solution frequencies. For our coding, one could speak of a consensus emerging regarding certain requirements for the written lesson plans in NRW and Berlin.

Subsequently, one-dimensional Rasch scaling was conducted with respect to T1 and T2 (T1 $n = 172$; T2 $n = 138$). Due to the extremely high solution frequency (100%) at T2, Item PN4P3 (*checking results and feedback*) was removed because it could not be used in the *ConQuest* analysis.

At both measurement time points, the scale proved to be reliable, with the items spreading sufficiently across the ability range (T1: $EAP = 0.75$; $\Theta = 1.25$). However, at the second time point, EAP reliability and theta variance were significantly lower (T2 2: $EAP = 0.41$; $\Theta = 0.31$), possibly because the empirical variance de-

Table 5 Items' solution frequencies at the beginning (*T1*) and end (*T2*) of the induction phase

Degrees of explication	Item	Facility (as a %)	t1	t2	Differences (t2-t1)	t-Test	p-value (Wilcoxon)
Naming	CN1	68.9	36.2	92.8	56.6	0.000	0.000
Naming	CN2	43.0	12.3	75.4	63.1	0.000	0.000
Reasoning	CR1	10.0	6.8	10.9	4.1	0.566	0.857
Reasoning	CR2	51.5	15.7	88.4	72.7	0.000	0.000
Linking	CL1	37.9	10.2	67.4	57.2	0.000	0.000
Linking	CL2	60.8	22.6	97.8	75.2	0.000	0.000
Linking	CL3	57.3	18.3	97.1	78.8	0.000	0.000
Linking	CL4	22.3	12.8	28.3	15.5	0.112	0.000
Naming	PN1	97.7	71.5	97.1	25.6	0.287	0.366
Naming	PN2P1	98.1	71.1	98.6	27.5	0.475	1
Naming	PN3P2	98.1	71.1	98.6	27.5	0.475	1
Naming	PN4P3	98.7	71.1	100	28.9	0.629	0.48
Naming	PN5P4	51.8	36.6	53.6	17	0.931	0.553
Reasoning	PR1P1	85.8	57.4	94.2	36.8	0.158	0.001
Reasoning	PR2P2	74.1	41.3	95.7	54.4	0.000	0.000
Reasoning	PR3P3	78.6	53.6	84.8	31.2	0.407	0.047
Reasoning	PR4P4	33.7	17.0	46.4	29.4	0.044	0.001
Linking	PL1P1P2	31.7	5.5	61.6	56.1	0.000	0.000
Linking	PL2P2P3	22.0	3.8	42.8	39	0.000	0.000
Linking	PL3P3P4	3.9	2.1	5.1	3	0.602	0.593
Linking	PL4P1P2	44.7	27.7	52.9	25.2	0.400	0.032
Linking	PL5P2P3	23.9	11.5	34.1	22.6	0.109	0.000
Linking	PL6P3P4	5.8	4.7	5.1	0.4	0.266	0.371

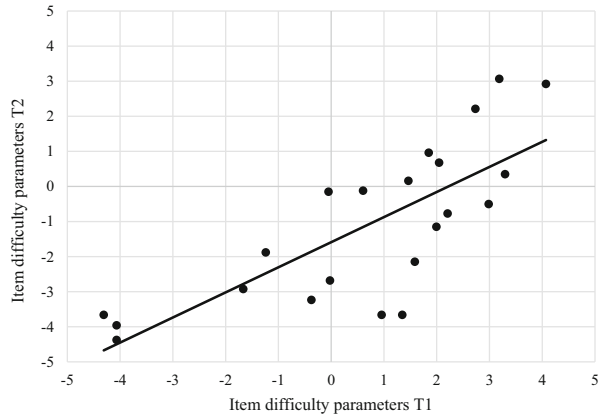
terminated at the second measurement time point was not very large. The low variances indicate a greater homogeneity in the data which could be interpreted as showing the effect of the training on the teacher' lesson planning competence.

A high correlation ($r=0.77$) exists between the item difficulties of the two measurement time points (59% common variance). The correlation illustrates that the item difficulties at both measurement time points were in a sufficiently invariant order for the two scales. We view this as an important indication of the measurement invariance (see Bond and Fox 2007). Fig. 2 provides the item difficulty parameter estimates in a bivariate scatter plot. Considering that the item parameters are quite close to the diagonal, we assumed sufficient sample independence in the tests.

The mean score¹ (total score) increased significantly ($p < 0.000$) from the beginning (*T1*: $M=9.36$, $SE=0.25$, $SD=3.3$) to the end (*t2*: $M=15.28$; $SE=0.19$, $SD=2.2$) of the induction. To determine whether an increase in measured plan-

¹ In the previous study, the WLEs from the different scales were not reported at that time, but the scores for the entire scale and the subscales were. To ensure comparability, we chose the same procedure for the presentation of results and compared the scores at the two measurement times. Under RQ 4, we refer to the WLEs to draw a more differentiated picture.

Fig. 2 Item difficulty parameters of the scales at both measurement time points



ning competence existed in relation to structuring, corresponding mean differences were calculated. Table 6’s second column contains the differences in the means at the two measurement time points. The mean differences of the total score and the two subscales are significant. The two columns on the right side of Table 7 indicate that these significant mean differences also were reflected in the corresponding effect sizes (*Cohen’s d* for different-size groups; T1 = 171, T2 = 138). Using the panel sample of pre-service teachers who submitted their lesson plans at T1 and T2 ($n = 116$), the mean differences were significant ($t[1,115] = -14.71, p < 0.000$) and practically relevant ($d = 2.05$). Furthermore, significant increases were found in all other areas for the panel sample.

The negative correlations could be interpreted as an indication of the induction’s efficacy because they demonstrate that the lesson plans about structuring as a planning aspect are significantly worse at the beginning of the induction than at the end.

In the next step, we examined the extent to which a change occurred in the different explication levels during the induction phase. Fig. 3 provides the distribution of explication levels at each time point. We used ability estimates from the three-dimensional IRT scaling in the virtual cases approach to indicate the three explica-

Table 6 Mean value differences and effect strengths

	<i>M</i> (difference)	<i>SD</i> (difference)	<i>Sig.</i> (2-sided)	<i>r</i>	<i>Cohens d</i>
TS_T1 vs. TS_T2	5.92	1.1	0.000	-0.13	2.07
SC_T1 vs. SC_T2	3.73	1.15	0.000	-0.16	2.15
SP_T1 vs. SP_T2	2.2	0.05	0.000	-0.03	1.17
SN_T1 vs. SN_T2	1.07	0.22	0.000	-0.11	1.26
SR_T1 vs. SR_T2	1.56	0.42	0.000	-0.07	1.27
SL_T1 vs. SL_T2	3.26	0.36	0.000	-0.01	1.95

TS Total Score, SC Score Contextualization, SP Score Phrasing, SN Score Naming, SR Score Reasoning, SL Score Linking

** $p < 0.01$, * $p < 0.05$

Table 7 Findings from regression analysis predicting planning skill scores at each time point

		Type 1	Type 2	(corrected) R^2
		β	β	
T1 ($n = 172$)	TotalScoreT1	-0.11	-0.45***	0.16
	Score_ContextualisationT1	-0.16*	-0.54***	0.23
	Score_PhasingT1	-0.01	-0.18*	0.02
	Score_NamingT1	-0.11	-0.32***	0.07
	Score_ReasoningT1	-0.01	-0.32***	0.09
	Score_LinkingT1	-0.14 ⁺	-0.43***	0.14
T2 ($n = 138$)	TotalScoreT2	0.01	0.09	-0.01
	Score_ContextualisationT2	0.27**	0.34***	0.08
	Score_PhasingT2	-0.13	-0.09	0.00
	Score_NamingT2	0.31***	0.72***	0.38
	Score_ReasoningT2	-0.14	0.08	0.03
	Score_LinkingT2	-0.07	-0.25*	0.03

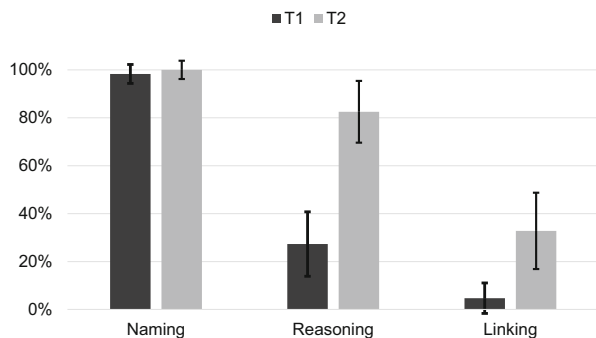
β standardised regression coefficient, T1 First written plan, T2 Last written plan

⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

tion levels in the written plans at T1 and T2. We used the item parameter estimates' mean values (Fig. 1) to create an ability summary at both time points. The black bars in Fig. 3 contained the results at T1 and the light gray bar at T2. Naming was at a significantly high level at both T1 and T2. The pre-service teachers made significant progress in reasoning and linking between T1 and T2. At the beginning of the internship, less than 30% of the instructional designs reached the justification level, and less than 5% of the instructional designs showed signs of linking decisions. By the end of induction, more than 80% of the lesson plans demonstrated reasoned decision making, and more than 30% demonstrated linking of these decisions.

This increase also is reflected in the mean differences and effect sizes, as provided in Table 6 (see above). Mean effect sizes are provided for the two lower levels. The results indicate that the ability to structure a lesson improves during the introductory phase of teacher education. Consequently, pre-service teachers can better justify and link their planning decisions at the end of the introductory phase (T2).

Fig. 3 Distribution of explication levels at each time point (with a 95% confidence interval)



4.4 RQ 4

Furthermore, we will examine whether the empirically found structure of structuring as an aspect of planning competence can be confirmed in different sub-samples (NRW and Berlin). In the previous sections, it was demonstrated that a statistically and practically significant increase takes place over time. However, a question arises: To what extent does this growth occur to a similar extent at the different locations (NRW and Berlin), depending on the pre-service teachers' teaching type? In view of the assumption that teacher education does have an impact, this may be expected, but it cannot be taken for granted due to the repeatedly mentioned problem of arbitrariness in teacher education curriculum and assessment (Strietholt and Terhart 2009).

Using a single-factor ANOVA for the *panel sample*, the overall mean difference between the three teaching types was found to be significant for the first time point ($p < 0.001$). A post-hoc test (Bonferroni) found a significant difference ($p < 0.001$) between Secondary Level I and Secondary Levels I and II in NRW ($p < 0.001$) and a significant difference between Secondary Levels I and II and the pre-service teachers in Berlin ($p < 0.000$). At the second measurement time point, no significant differences were found between the three teaching types.

To obtain a better insight into the development of structuring as part of planning competence, we conducted a regression analysis for each time point. We recorded planning skills as a dependent variable and teaching type as an independent variable. Teaching type was specified using dummy variables: Type 1 was coded as the first dichotomous variable and included the type "lower secondary level" (coded as 1). As the second dichotomous variable, Type 2 recorded the type "Secondary Levels I and II." The pre-service teachers in Berlin served as a reference group. At the first time point, significant differences were found between teaching types, and 16% (corrected R^2) of the differences between the planning performance (total score) based on the teaching type can be explained. At the second time point, the training locations did not differ, indicating a high degree of agreement in teachers' lesson plans (approximately 1% in differences can be explained; corrected R^2).

The findings reveal that at T1, the quality of teacher's lesson plans depend on their teaching type, but hardly any differences exist between teaching types and teacher training at both locations. Teaching types contributed to the fact that significant growth exists regarding lesson planning under the aspect of structuring.

5 Summary and discussion

Only few studies on modelling and measuring planning competence have been conducted so far. We used a standardized method for assessing written lesson plans developed in a previous study. To verify and replicate the findings, we conducted the present scaling-up study and examined 310 written lesson plans at two measurement time points from pre-service teachers during induction in NRW and Berlin.

The findings from the previous study were confirmed. At the content level, two subscales (*contextualization* and *phasing*) could be identified. The Chi-square test

provides a better fit for the two-dimensional modelling than the one-dimensional (RQ 1). Furthermore, three different explication levels for planning decisions (naming, reasoning, and linking) were theoretically derived and empirically proven, which also showed a better fit compared with the one-dimensional modelling (RQ 2). At the end of the induction, the pre-service teachers showed a higher degree of planning competence. Measures of lesson-structure planning increased during induction significantly (RQ 3) in all three teaching types (RQ 4). However, the increase was higher in NRW, although performance values at the second time point did not differ significantly between the three groups.

Overall, the results indicate that the present test instrument reflects existing content in teacher education across federal state boundaries. This indicates that the construct has high *curricular validity*, which is oriented closely toward the induction context and provides evidence that lesson planning in general and structuring specifically are central elements of teacher training. The significant increase in instructional planning competence also demonstrates the effectiveness of teacher induction. The pre-service teachers succeeded in better justifying and linking their instructional planning decisions. Therefore, it can be assumed that the category inventory used captures content that is curricular relevant during induction.

However, this also raises the question: To what extent do institutional requirements determine pre-service teachers' planning decisions? So, the evaluation procedure could be criticized because the lesson plans are essential part of the examination and certification processes and, thus, may contain teacher educators' individual preferences regarding demonstration lessons. Therefore, to gather further evidence on content validity, it would be necessary to ask teacher educators whether the system of categories established here actually covers what is taught in everyday educational contexts and what requirements they view as important in the context of initial teacher education. This also would provide evidence of the transparency of the assessment criteria applied in teacher induction, as Strietholt and Terhart (2009) found that only about 50% of teacher educators in Germany make their assessment criteria sufficiently transparent to pre-service teachers.

A possible interpretation is that these increases relate to the examination procedure during induction. The increase could just be the consequence of the intensive preparation for the final examination. Surely a deficit exists in (*quasi-*)*experimental studies* that deal with the (causal) analysis of the development of lesson-planning competence during induction. So far, learning opportunities have hardly been investigated (König et al. 2017). It is unclear how pre-service teachers receive content input and/or gain practical experience during induction that help them develop their planning competences.

The findings from the previous study and the replication study can be used to develop quality indicators for assessment practices in teacher education. However, the category system is very extensive. It could be examined whether this category system can be made more efficient and whether individual categories might need to be combined. It would be conceivable that categories S4 and S5, which record the linking of the phase decisions both in the text and in the articulation scheme, represent one category. A teacher who has linked the individual phases argumentatively

in the text could do without such explanations in the articulation scheme to “inflate” the lesson plan too much.

Furthermore, the question arises as to whether the “contextualization” and “phasing” subscales together form one scale (under the aspect of structuring) or whether they do not represent two separate constructs that are parts of pre-service teachers’ planning competence. Information concerning the complexity of the construct of planning competence was provided by König et al. (2021). In their *CODE-PLAN* model (*cognitive demands of lesson planning*), six factors (content transformation, task creation, adaptation to student learning dispositions, clarity of learning objectives, unit contextualization, and phasing) were combined as indicators of planning competence. The two subscales used here were included in the analysis as separate constructs. Taking the model into account, the six different cognitive demands on lesson planning can be combined into an overall construct of lesson planning competence. This model would be empirically founded and could be used as a criterion variable at the end of the induction phase.

Lesson planning is viewed as being important for the teacher’s performance in the classroom, presumably influencing the procedural flow of lesson design and students’ learning outcomes. These assumptions are plausible, but so far hardly supported by evidence from empirical research (Besa et al. 2020; Rothland 2021). Currently, there is the need for research on the relationship between lesson planning measures and the characteristics of process and outcome quality in teaching. From a practical perspective, the question is whether a good plan can predict performance in the classroom (Shavelson 1987). The existence of a good plan does not guarantee that instruction actually will proceed in a well-structured way. The teaching process is too complex and unpredictable for that (Doyle 2006). Therefore, further research is needed to apply the aspect of structuring not only to the planning situation, but also to the actual teaching process. For example, an important question is whether the planned structuring measures contribute to smoother teaching. This also would relate to effective classroom management. Previous studies on classroom management have focused on planning teacher behavior in advance (Doyle 2006; Everson and Weinstein 2006). The present construct could be used to test whether well-structured instruction positively impacts effective classroom management and positively influences student learning. To gather evidence for prognostic validity, written plans for demonstration lessons need to be collected from pre-service teachers, and other lessons also need to be assessed. Students could conduct these assessments, or they could be conducted through video recordings of lessons or by trained observers.

Supplementary Information The online version of this article (<https://doi.org/10.1007/s11618-022-01125-9>) contains supplementary material, which is available to authorized users.

Funding This work was supported by the Federal Ministry of Education and Research, Germany [Bundesministerium für Bildung und Forschung, BMBF, grant number 01PK15014A, 01PK15014B, 01PK15014C].

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest M. Krepf and J. König declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris: OECD.
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: an approach to errors in variables regression. *Journal of educational and behavioral Statistics*, 22(1), 47–76. <https://doi.org/10.3102/10769986022001047>.
- Arlin, M. (1979). Teacher transitions can disrupt time flow in classrooms. *American Educational Research Journal*, 16(1), 42–56. <https://doi.org/10.3102/00028312016001042>.
- Arnold, K. H., & Koch-Priewe, B. (2011). The merging and the future of the classical German traditions in general didactics: a comprehensive framework for lesson planning. In B. Hudson & M. A. Meyer (Eds.), *Beyond fragmentation: didactics, learning and teaching in europe* (pp. 252–264). Opladen: Barbara Budrich.
- de Ayala, R. J., Schafer, W. D., & Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *British Journal of Mathematical and Statistical Psychology*, 48(2), 385–405. <https://doi.org/10.1111/j.2044-8317.1995.tb01070.x>.
- Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9(4), 469–520. <https://doi.org/10.1007/s11618-006-0165-2>.
- Berliner, D. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society*, 24, 200–212. <https://doi.org/10.1177/0270467604265535>.
- Besa, K.-S., Gesang, J., & Hinzke, J.-H. (2020). Zum Verhältnis von Forschungskompetenz und Unterrichtsplanung. In M. Keller-Schneider, K. Krammer & M. Trautmann (Eds.), *Jahrbuch für Allgemeine Didaktik 2020* (pp. 97–113). Hohengehren: Schneider Verlag.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies: viewing competence as a continuum. *Zeitschrift für Psychologie*, 223, 3–13. <https://doi.org/10.1027/2151-2604/a000194>.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd edn.). Mahwah: Erlbaum.
- Borich, G. D. (2004). *Effective teaching methods* (5th edn.). Upper Saddle River: Merrill Prentice Hall.
- Brophy, J. (2000). *Teaching* (Educational practices series, Vol. 1). Brüssel: International academy of education. <http://www.ibe.unesco.org>
- Calderhead, J. (1996). Teachers beliefs and knowledge. In D. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 709–725). New York: Simon and Schuster, Macmillan.
- Cochran-Smith, M., & Villegas, A. M. (2016). Research on teacher preparation: charting the landscape of a sprawling field. In D. H. Gitomer & C. A. Bells (Eds.), *Handbook of research on teaching* (5th edn., pp. 439–547). Washington, DC: AERA.
- von Davier, A. A., Carstensen, C. H., & von Davier, M. (2006). Linking competencies in horizontal, vertical, and longitudinal settings and measuring growth. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 53–80). Göttingen: Hogrefe.
- Doenau, S. J. (1987). Structuring. In M. J. Dunkin (Ed.), *The international encyclopedia of teaching and teacher education* (pp. 398–407). New York: Pergamon Press.
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5th edn.). Berlin, Heidelberg: Springer.
- Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice, and contemporary issues* (pp. 107–136). New York: Routledge.
- Drollinger-Vetter, B. (2011). *Verstehenselemente und Strukturelle Klarheit. Fachdidaktische Qualität der Anleitung von Mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann.

- Esslinger-Hinz, I., Wigbers, M., Giovannini, N., Hannig, J., Herbert, L., Jäkel, L., Klingmüller, C., Lange, B., Neubrech, N., & Schnepf-Rimsa, E. (2013). *Der ausführliche Unterrichtsentwurf*. Weinheim: Beltz.
- Evertson, C. M., & Weinstein, C. S. (2006). Classroom management as a field of inquiry. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: research, practice, and contemporary issues* (pp. 3–15). New York: Routledge.
- Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613–619. <https://doi.org/10.1177/001316447303300309>.
- Good, T. L., & Brophy, J. E. (2007). *Looking in classrooms*. Boston: Allyn & Bacon.
- Grzesik, J. (1979). *Unterrichtsplanning*. Heidelberg: Quelle & Meyer.
- Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. Abingdon: Routledge.
- Heimann, P. (1972). Didaktik 1965. In P. Heimann, G. Otto & W. Schulz (Eds.), *Unterricht. Analyse und Planung* (6th edn., pp. 7–12). Hannover: Schroedel.
- Hiebert, J., Gallimore, R., Garnier, H., Giwin, K. B., Hollingsworth, H., Jacobs, J., Chui, A. M.-Y., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., & Stigler, J. (2003). *Teaching mathematics in seven countries: results from the TIMSS 1999 video study*. Washington, DC: U.S. Department of Education.
- Hohenstein, F., Zimmermann, F., Kleickmann, T., Köller, O., & Möller, J. (2014). Sind die bildungswissenschaftlichen Standards für die Lehramtsausbildung in den Curricula der Hochschulen angekommen? *Zeitschrift für Erziehungswissenschaft*, 17(4), 497–507. <https://doi.org/10.1007/s11618-014-0563-9>.
- Jäger, S., & Maier, U. (2019). Unterrichtsplanning. In E. Kiel, B. Herzig, U. Maier & U. Sandfuchs (Eds.), *Handbuch Unterrichten an allgemeinbildenden Schulen* (pp. 455–466). Bad Heilbrunn: Klinkhardt.
- John, P. D. (2006). Lesson planning and the student teacher: Re-thinking the dominant model. *Journal of Curriculum Studies*, 38(4), 483–498. <https://doi.org/10.1080/00220270500363620>.
- Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research*, 62(2), 129–169. <https://doi.org/10.2307/1170578>.
- Kaiser, A., & Kaiser, R. (1991). *Studienbuch Pädagogik: Grund- und Prüfungswissen* (5th edn.). Frankfurt a. M.: Cornelsen Scriptor.
- Klafki, W. (1995). Didactic analysis as the core of preparation of instruction. *Journal of Curriculum Studies*, 27(1), 13–30. <https://doi.org/10.1080/0022027950270103>.
- Klafki, W. (2007). *Neue Studien zur Bildungstheorie und Didaktik. Zeitgemäße Allgemeinbildung und kritisch-konstruktive Didaktik* (6th edn.). Weinheim: Beltz.
- Kleickmann, T. (2012). *Kognitiv aktivieren und inhaltlich strukturieren im naturwissenschaftlichen Sachunterricht* (Publikation des Programms SINUS an Grundschulen). Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN).
- Klingberg, L. (1982). *Unterrichtsprozesse und didaktische Fragestellung: Studien und Versuche*. Berlin: Volk und Wissen.
- König, J. (2012). Die Entwicklung von pädagogischem Unterrichtswissen: Theoretischer Rahmen, Testinstrument, Skalierung und Ergebnisse. In J. König & A. Seifert (Eds.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (pp. 143–182). Münster: Waxmann.
- König, J., & Blömeke, S. (2009a). Pädagogisches Wissen von österreichischen Lehramtsstudierenden. *Erziehung & Unterricht*, 159(1/2), 175–186.
- König, J., & Blömeke, S. (2009b). Pädagogisches Wissen von angehenden Lehrkräften: Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerausbildung. *Zeitschrift für Erziehungswissenschaft*, 12(3), 499–527. <https://doi.org/10.1007/s11618-009-0085-z>.
- König, J., Buchholtz, C., & Dohmen, D. (2015). Analyse von schriftlichen Unterrichtsplanningen. Empirische Befunde zur didaktischen Adaptivität als Aspekt der Planungskompetenz angehender Lehrkräfte. *Zeitschrift für Erziehungswissenschaft*, 18, 375–404. <https://doi.org/10.1007/s11618-015-0625-7>.
- König, J., Bremerich-Vos, A., Buchholtz, C., Lammerding, S., Strauß, S., Fladung, I., & Schleiffer, C. (2017). Die Bedeutung des Professionswissen von Referendarinnen und Referendaren mit Fach Deutsch für ihre Planungskompetenz (PlanvoLL-D). In S. Wernke & K. Zierer (Eds.), *Die Unterrichtsplanung: Ein in Vergessenheit geratener Kompetenzbereich?! Status Quo und Perspektiven aus Sicht der empirischen Forschung* (pp. 121–133). Bad Heilbrunn: Klinkhardt.

- König, J., Bremerich-Vos, A., Buchholtz, C., Fladung, I., & Glutsch, N. (2020a). Pre-service teachers' generic and subject-specific lesson-planning skills. On learning adaptive teaching during initial teacher education. *European Journal of Teacher Education*, 43(2), 131–150. <https://doi.org/10.1080/02619768.2019.1679115>.
- König, J., Bremerich-Vos, A., Buchholtz, C., & Glutsch, N. (2020b). General pedagogical knowledge, pedagogical adaptivity in written lesson plans, and instructional practice among preservice teachers. *Journal of Curriculum Studies*, 52(6), 800–822. <https://doi.org/10.1080/00220272.2020.1752804>.
- König, J., Krepf, M., Bremerich-Vos, A., & Buchholtz, C. (2021). Meeting cognitive demands of lesson planning. Introducing the CODE-PLAN model to describe and analyze teachers' planning competence. *The Teacher Educator*, 56(4), 466–487. <https://doi.org/10.1080/08878730.2021.1938324>.
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart & Winston.
- Krepf, M., & König, J. (under review). Structuring the lesson: an empirical investigation of pre-service teacher decision-making during the planning of a demonstration lesson. *Journal of Education for Teaching*.
- Kyriakides, L., Creemers, B. P. M., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: the contribution of the dynamic model. *ZDM-Mathematics Education*, 50, 381–393. <https://doi.org/10.1007/s11858-018-0919-3>.
- Leinhardt, G., & Greeno, J. (1986). The cognitive kill of teaching. *Journal of Educational Psychology*, 78(2), 75–95.
- Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (2nd edn., pp. 69–105). Berlin: Springer.
- Maier, U. (2012). *Lehr-Lernprozesse in der Schule: Studium*. Bad Heilbrunn: Klinkhardt.
- Maulana, R., Opendakker, M.-C., Stroet, K., & Bosker, R. (2012). Observed lesson structure during the first year of secondary education: exploration of change and link with academic engagement. *Teaching and Teacher Education*, 28, 835–850. <https://doi.org/10.1016/j.tate.2012.03.005>.
- Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution. Retrieved September 12, 2022 from <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173>
- Meyer, H. (1999). *Leitfaden zur Unterrichtsvorbereitung*. Frankfurt a. M.: Cornelsen Scriptor.
- Meyer, H. (2004). *Was ist guter Unterricht?* Berlin: Cornelsen Scriptor.
- Muijs, D., & Reynolds, D. (2011). *Effective teaching: evidence and practice* (3rd edn.). London: SAGE.
- Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning. *School effectiveness and school improvement*, 25(2), 231–256. <https://doi.org/10.1080/09243453.2014.885451>.
- Pauli, C., & Reusser, K. (2003). Unterrichtsskripts im schweizerischen und deutschen Mathematikunterricht. *Unterrichtswissenschaft*, 31(3), 238–272.
- Plöger, W. (2008). *Unterrichtsplanung: Ein Lehr- und Arbeitsbuch für Studium und Seminar*. Köln: Kölner Studien Verlag.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.
- Rey, T., Lohse-Bossenz, H., Wacker, A., & Heyl, V. (2018). Adaptive Planungskompetenz bei angehenden Lehrkräften in der zweiten Phase der Lehrerbildung: Befunde einer Pilotierungsstudie aus Baden-Württemberg. *HEI Education Journal*, 1(1/2), 127–150. <https://doi.org/10.17885/heiup.heied.2018-1-2.23829>.
- Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 376–391). New York: MacMillan.
- Rost, D. H. (2007). *Interpretation und Bewertung pädagogisch-psychologischer Studien* (2nd edn.). Weinheim: Beltz.
- Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.
- Rothland, M. (2021). Anmerkungen zur Modellierung und Operationalisierung (allgemeindidaktischer) Unterrichtsplanungskompetenz. *Unterrichtswissenschaft*. <https://doi.org/10.1007/s42010-021-00111-0>.
- Scholl, D., & Plöger, W. (2020). Unterricht als Gegenstand der Lehrerbildung. In C. Cramer, J. König, M. Rothland & S. Blömeke (Eds.), *Handbuch Lehrerinnen- und Lehrerbildung* (pp. 21–30). Bad Heilbrunn: Klinkhardt.
- 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. <https://doi.org/10.3102/0034654307310317>.

- Seidel, T., Rimmele, R., & Prenzel, M. (2005). Clarity and coherence of lesson goals as a scaffold for student learning. *Learning and Instruction*, *15*, 539–556. <https://doi.org/10.1016/j.learninstruc.2005.08.004>.
- Shavelson, R.J. (1983). Review of research on teachers' pedagogical judgments, plans, and decisions. *Elementary School Journal*, *83*(4), 392–413. <https://doi.org/10.1086/461323>.
- Shavelson, R.J. (1987). Planning. In M.J. Dunkin (Ed.), *The international encyclopedia of teaching and teacher education* (pp. 483–486). New York: Pergamon Press.
- Shavelson, R.J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgements, decisions and behavior. *Review of Educational Research*, *51*(4), 455–498. <https://doi.org/10.3102/00346543051004455>.
- Shulman, L.S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. <https://doi.org/10.3102/0013189X015002004>.
- Shulman, L.S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Research*, *57*(1), 1–22. <https://doi.org/10.17763/haer.57.1.j463w79r56455411>.
- Strietholt, R., & Terhart, E. (2009). Referendare beurteilen. Eine explorative Analyse von Beurteilungsinstrumenten in der zweiten Phase der Lehrerbildung. *Zeitschrift für Pädagogik*, *55*(4), 622–645. <https://doi.org/10.25656/01:4266>.
- Walberg, H.J., & Paik, S.J. (2000). *Effective educational practices* (Educational practices series, Vol. 3). Brüssel: International academy of education. <http://www.ibe.unesco.org>
- Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, *54*, 427–450. <https://doi.org/10.1007/BF02294627>.
- Westermann, D.A. (1991). Expert and novice teacher decision making. *Journal of Teacher Education*, *42*(4), 292–305. <https://doi.org/10.1177/002248719104200407>.
- Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.
- Wu, M., & Adams, R.J. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, *18*(2), 93–113. <https://doi.org/10.1007/BF03217438>.
- Wu, M., Adams, R.J., & Wilson, M. (1997). *Conquest: multi-aspect test software* (computer program). Camberwell: Australian council for educational research.
- Yinger, R.J. (1980). A study of teacher planning. *Elementary School Journal*, *80*(3), 107–127. <https://doi.org/10.1086/461181>.
- Zahorik, J.A. (1970). The effect of planning on teaching. *Elementary School Journal*, *71*(3), 143–151. <https://doi.org/10.1086/460625>.