Received: 20 December 2022 | Revised: 22 June 2023 | Accepted: 12 August 2023

DOI: 10.1002/sce.21832

RESEARCH ARTICLE

Science Education WILEY

Assessing the quality of science teachers' lesson plans: Evaluation and application of a novel instrument

Leroy Großmann | Dirk Krüger

Biology Education, Freie Universität Berlin, Berlin, Germany

Correspondence
Leroy Großmann, Biology Education, Freie Universität Berlin, Schwendenerstraße 1, Berlin 14195, Germany
Email: leroy.grossmann@fu-berlin.de

Abstract

Lesson planning is a core part of teachers' professional competence. Written lesson plans play a significant role in science teacher education as a preparation for demonstration lessons during the final teacher certification exam. However, the few existing scoring rubrics on lesson plans are not particularly theoretically sound and are barely tested for the validity of score interpretations. In response to the demand for transparent and applicable criteria, we developed the *rubric to assess science lesson plans* (RALP) to assess science teachers' lesson plan quality. We employed a mixed-methods approach: First, we present multiple sources of validity evidence (based on *test content*, *internal structure*, *relations to other variables*, and *consequences of testing*) as mainly quantitative indicators for the quality of the RALP. Based upon that, we applied the RALP to lesson plans written by preservice and trainee science teachers ($N = 100$) and provided a qualitative analysis of six cases to illustrate common patterns in these lesson plans. Results indicate that teacher educators consider the RALP criteria ($N = 24$) relevant and objectively applicable. Correlation analyses of the scores and two teacher educators' holistic quality assessment of all lesson plans provide convincing evidence that the RALP can discriminate lesson plan quality

levels. Moreover, comparisons between preservice science teachers and trainee science teachers reveal that trainee teachers score significantly higher than preservice teachers, indicating that the RALP is sensitive to differences in teaching and planning experience. The application and in-depth analysis of three criteria of the RALP illustrate these differences in levels of planning quality. We discuss possible applications of the RALP in science teacher education and research in science teaching.

KEYWORDS

assessment, lesson plan, science teacher education, scoring rubric, validity

1 | INTRODUCTION

Lesson planning is a core part of teachers' professional competence (Carlson et al., 2019; Mutton et al., 2011; Zaragoza et al., 2021). Even though empirical evidence is still scarce, it is plausible to assume that high-quality planning correlates with high-quality classroom teaching. Lesson planning was shown to predict teaching effectiveness in terms of students' achievement gains (Darling-Hammond et al., 2013) and in terms of students' ratings of instructional quality (König et al., 2021). Consequently, lesson planning forms a constitutive part of teacher education, both in national professional standards (e.g., Australian Institute for Teaching and School Leadership, 2018; France: Ministère de l'Éducation Nationale, 2013; Germany: Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK], 2019) and in official teacher accreditation requirements such as the American "educative Teacher Performance Assessment" (Sato, 2014). Thus, written lesson plans play a significant role in teacher training (König et al., 2020). In a lesson plan, teachers describe and justify a planned instructional setting. Prospective teachers write such lesson plans to get feedback for their planning, teaching, and reflection skills from their teacher educators who visit them at school and observe a planned lesson. Consequently, lesson planning, in general, and writing proper lesson plans, in particular, are key competencies prospective teachers are intended to acquire during teacher education (Morine-Dershimer, 2011).

However, empirical research on lesson planning as a situation-specific skill of (prospective) science teachers is still scarce (Campbell et al., 2022; König et al., 2021). Krepf and König (2022, p. 14) argue that "there are hardly any scientifically proven criteria for the analysis and evaluation of written lesson plans, therefore [we] consider it an important task of empirical teacher education research to present suitable procedures for the assessment of written lesson plans." Specifically, Kang (2017) points out that "research on teachers' planning was often conducted in a controlled setting to uncover cognitive decisions made by teachers. Researchers tend to rely on either teacher self-reported data or data generated from think-aloud methods […]. Little empirical evidence focuses on teachers' effective planning in the natural teaching environments" (Kang, 2017, p. 56). Scoring rubrics can be an evidentially helpful and effective instrument for assessing the quality of written plans. They are well-established in many research areas, such as education (Jonsson & Svingby, 2007) and higher education (Brookhart, 2018). Regarding research on lesson planning, there is a need for an instrument that was developed in an ecologically valid manner and enables researchers in science education to assess the quality of prospective science teachers' lesson plans. To close this gap, the *rubric to assess science lesson plans* (RALP) was developed within the framework of pedagogical content knowledge (PCK; Carlson et al., 2019; Park & Oliver, 2008). However, the RALP would only add a fruitful

methodology to teacher education and research in science education if it was carefully evaluated to ensure that the inferences made from the scoring results are valid. In recent years, there has been an increasing demand for empirical research that devotes more attention to the consideration of validity issues generally in research (AERA, APA, & NCME, 2014; Kane, 2013; Pellegrino et al., 2016) and particularly regarding scoring rubrics (Brookhart, 2018; Moskal & Leydens, 2000). However, this is not the case in many studies that present scoring rubrics in higher education (Reddy & Andrade, 2010), PCK research (Chan et al., 2019), or research on lesson planning. Hence, in this study, we aim to employ a mixed-methods approach (Creswell & Plano Clark, 2018) to evaluate the quality of the RALP concerning multiple sources of validity evidence for the interpretations of RALP scores. Specifically, we build a validity argument (Kane, 2013) referring to validity evidence based on *test content*, *internal structure*, *relations to other variables*, and *consequences of testing* (AERA, APA, & NCME, 2014). Based upon that, we will apply the RALP to original lesson plans and provide qualitative insights into prospective biology teachers' lesson planning competence.

2 | LESSON PLANNING AT THE HEART OF TEACHERS' PROFESSIONAL COMPETENCE

Nearly four decades have passed since Shulman (1986) introduced the differentiation of content knowledge (CK), pedagogical knowledge (PK), and, first and foremost, PCK as the unique component of teachers' professional knowledge. As far as lesson planning is concerned, only a few empirical studies explicitly locate themselves within Shulman's (1986) topology of teachers' professional knowledge or frameworks building upon that, for example, the *refined consensus model* (RCM) of PCK (Carlson et al., 2019). PCK is defined as the "knowledge of, reasoning behind, and planning for teaching a particular topic in a particular way for a particular purpose to particular students for enhanced student outcomes" (Gess-Newsome, 2015, p. 36). Carlson et al. (2019) recognize that PK, CK, knowledge of students, knowledge of curriculum, and knowledge of assessment are foundational to PCK. Without these knowledge bases, a teacher's PCK is limited. The RCM was a significant step forward in PCK research because it combines the static knowledge bases (i.e., knowing *that*) and the dynamic process of enacting the knowledge in the teaching cycle of planning, teaching, and reflecting (i.e., knowing *how*). In the current paper, we connect to this conceptualization of PCK and explicitly focus on it when enacted in the lesson planning process (i.e., enacted PCK during planning, $ePCK_P$; Alonzo et al., 2019). Enacted PCK is part of an individual's personal PCK (pPCK). More specifically, we adapt the pentagon model of PCK (Park & Oliver, 2008) that describes the integration of some of the knowledge bases named in the outermost circle of the RCM. It consists of *orientations to teaching science* (OTS), *knowledge of the science curriculum* (KSC), *knowledge of students' understanding in science* (KSU), *knowledge of instructional strategies* (KISR), and *knowledge of assessment strategies* (KAs). Park and Oliver (2008) emphasize the importance of enacting PCK (i.e., planning, reflecting, and particularly teaching) and integrating the five components:

> This integration is accomplished through the complementary and ongoing readjustment by both reflection-in-action and reflection-on-action, resulting in strengthened coherence among the components […]. This model suggests that the development of one component within PCK will, in turn, influence the development of others, and ultimately enhance this holistic PCK. Because PCK, which comprises effective teaching, requires the integration of the components in highly complex ways, lack of coherence among components can be problematic in developing PCK and increased knowledge of a single component may not be sufficient to stimulate significant change in practice. (Park & Oliver, 2008, p. 814)

The interplay of reflection-in-action (i.e., teaching) and reflection-on-action (i.e., planning the lesson, and reflecting on it afterward) leads to an ongoing readjustment of the five PCK components. This process is

conceptualized as pedagogical reasoning (Park & Suh, 2019), which, in turn, forms the critical practice of the inner circle of the RCM (Carlson et al., 2019).

PCK is regarded to be the crucial knowledge base for teachers (Chan & Hume, 2019) due to two reasons: First, PCK is positively correlated with teaching quality in general (Park et al., 2011; Reynolds & Park, 2021), and with particular subaspects such as cognitive activation (Keller et al., 2017) or student learning support (Kunter et al., 2013). Second, teachers' level of PCK seems to predict their students' achievement (e.g., Förtsch et al., 2016; Kunter et al., 2013; Mahler et al., 2017).[1]

"[P]lanning is essentially knowledge-based" (Mutton et al., 2011, p. 412). When planning a lesson, teachers transform their professional knowledge, motivation, and beliefs into a (mental) lesson plan (Stender et al., 2017). Afterward, they consider the intended learning outcome, students' needs, and other relevant aspects related to classroom teaching and thus, modify their initial lesson plan. Prospective teachers need help to use their professional knowledge to justify planning decisions (Zaragoza et al., 2021). They use PCK in appropriate breadth and depth or neglect important aspects (Koberstein-Schwarz & Meisert, 2022). König et al. (2020) showed that trainee teachers teaching the school subject German elaborate on their planning decisions in lesson plans significantly more often by use of generic (PK) instead of subject-specific (PCK) features. Vogelsang et al. (2022) found a significant positive correlation between preservice physics teachers; PCK and their lesson planning competence (measured using a standardized performance assessment instrument) before and after a 6-month practical semester. In line with that, Backfisch et al. (2020) showed that lesson plans written by trainee and in-service mathematics teachers are of higher instructional quality than those written by preservice teachers. The instructional quality of these planned lessons correlated moderately with mathematics teachers' PCK.

In response to the need for common ground in empirical research on lesson planning, König et al. (2021) developed a heuristic model of the cognitive demands of lesson planning (CODE-PLAN model). According to that, teachers are confronted with six significant demands in the lesson planning process: *content transformation*, *task creation*, *adaptation to student learning dispositions*, *clarity of learning objectives*, *unit contextualization*, and *phasing*. These demands cover many components of PCK, for instance, as described in the pentagon model (Park & Oliver, 2008) that strengthens the necessity to interconnect the different components of PCK in planning, teaching, and reflecting. Only KAs are missing in the CODE-PLAN model, which was developed by a team of German researchers, possibly because the assessment of learning is widespread in the American educational system (e.g., Wiggins & McTighe, 2005) but barely occurs in German theoretical literature on lesson planning and advice literature for teachers (Vogelsang & Riese, 2017).

Nevertheless, it is reasonable to assume that a high level of PCK facilitates meeting the six abovementioned cognitive demands (König et al., 2021) to plan effective lessons. Both research on lesson planning and PCK have shown that, unlike experts, novices struggle to interconnect their professional knowledge (Westerman, 1991) and hence have more difficulties planning effective lessons (Koberstein-Schwarz & Meisert, 2022). In addition, Weitzel and Blank (2020) found that preservice biology teachers often do not consider certain aspects of lesson plans (e.g., students' conceptions, assessment of science learning, and the structure of teacher-class dialogs). One recently suggested way to foster students' ability to apply professional knowledge in lesson planning is scaffolding, which draws on research on professional vision (Zaragoza et al., 2023). Similar to this approach, we suggest providing a scoring rubric as another possibility of helping prospective teachers to meet those cognitive demands in lesson planning and interconnect four of the five PCK components.

3 | ASSESSING THE QUALITY OF LESSON PLANS

3.1 | The RALP

The RALP (Großmann & Krüger, 2022a, 2023) was developed based on the pentagon model of PCK (Park & Oliver, 2008) which served as a conceptual framework for our study. It illustrates the need to interconnect different

PCK components, which is necessary to overcome the challenges in the lesson planning process mentioned above. This study focuses on the four components that address knowledge (KSC, KAs, KISR, KSU). Although OTS is another crucial component of PCK as it shapes and is shaped by the remaining four components, we did not include it in the RALP development process. We aim to provide a scoring rubric that reflects theoretically and empirically based criteria to assess lesson plan quality rather than including science teachers' subjective beliefs about the purposes of learning science or the nature of science.

Since sophisticated PCK is expressed by the degree of interconnectedness of the PCK components (e.g., Chan & Hume, 2019) and since a high level of PCK correlates to teaching quality (Keller et al., 2017; Park et al., 2011; Reynolds & Park, 2021) and students' achievement (Förtsch et al., 2016; Kunter et al., 2013; Mahler et al., 2017), we developed criteria that explicitly target the interconnections between PCK components (Figure 1).

Twenty-four cognitive demands were identified based on König et al. (2021) that can be assigned to four of the five components of PCK. The 24 criteria were deduced from empirical and theoretical literature on teaching quality and from an extensive qualitative analysis of trainee science teachers' (TSTs') lesson plans (Großmann & Krüger, 2020, 2022c) using the PCK map approach (Park & Suh, 2019). Hence, the RALP reflects the realms of PCK needed in lesson planning: science teachers' pPCK and their $ePCK_p$ (Alonzo et al., 2019; Carlson et al., 2019). To give a concrete example of how the criteria were developed, criterion I4 (suitability of tasks) may provide an insight into the procedure: Creating tasks is a significant cognitive demand in lesson planning (König et al., 2021) and can be assigned to KISRs (Park & Oliver, 2008). One major challenge for prospective teachers is planning intellectually challenging tasks that directly relate to a specific intended learning outcome (Kang, 2017). Formulating intended learning outcomes to foster students' abilities is crucial. Criterion I4 explicitly builds a connection between these two aspects. Planning a lesson with tasks that do not relate to the intended learning outcome would correspond to level 0 in the RALP. Planning a task that addresses the intended learning outcome and analyzing the task in the lesson plan would correspond to level 1, and providing justifications for the design of the tasks (e.g., by relating to the students' current needs) would correspond to level 2. This procedure applies to all criteria assigned to the arrows in Figure 1.

To avoid the pitfalls of extant scoring rubrics in higher education, specific guidelines (Andrade, 2005; Brookhart, 2018; Dawson, 2017; Jonsson & Svingby, 2007), such as the use of descriptive language, have been

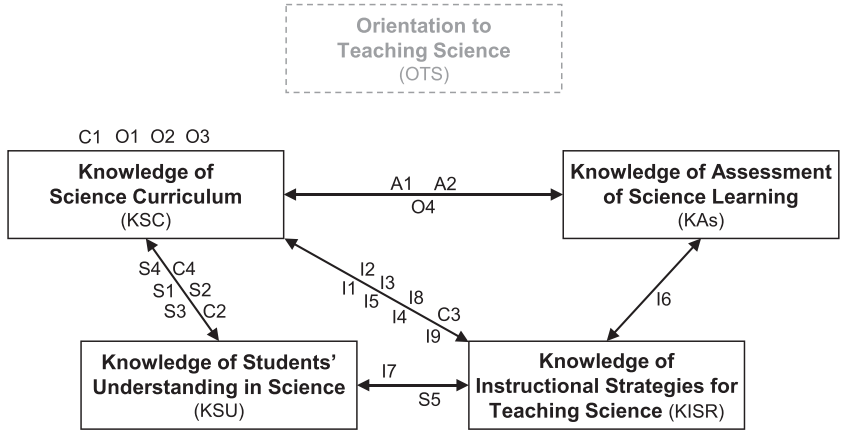

FIGURE 1 Assignment of the 24 RALP criteria to the components and their interconnections in the pentagon model of PCK (Park & Oliver, 2008). The relative closeness of criteria on the interconnecting arrows indicates which of the two components the criterion focuses on most. For this study, the PCK component "Orientation to Teaching Science" was not considered and is therefore grayed out. PCK, pedagogical content knowledge; RALP, rubric to assess science lesson plans.

considered. The 24 RALP criteria were deduced from an extensive qualitative analysis of TSTs' lesson plans (Großmann & Krüger, 2022c) using the PCK map approach (Park & Suh, 2019). The comparison of lesson plans with highly and poorly interconnected PCK led to identifying three performance-level descriptions for all criteria. Connecting to König et al.'s (2021), six cognitive demands in lesson planning, we were able to add and specify another 18 (Großmann & Krüger, 2023), adding up to a total of 24 RALP criteria (Supporting Information: Appendix A). For instance, I4 refers to the tasks that students will be confronted with, which should be well-aligned with the intended learning outcome. Hence, I4 is positioned on the interconnection between KISR and OTS. Except for O1, O2, O3, C1, and A2, all criteria emphasize the need to interconnect PCK components (Figure 1). An extensive description of the development of the RALP is provided by Großmann and Krüger (2023).

3.2 | How the RALP can be beneficial for science teacher education

Rubrics can be used for research purposes on the one hand and for formative or summative assessment in teacher education on the other hand (Krebs et al., 2022; Panadero et al., 2018; Panadero & Jonsson, 2013). This would also account for lesson planning (Ozogul et al., 2008), which is a crucial aspect to teach in science teacher training (Drost & Levine, 2015; Karlström & Hamza, 2021). Since lesson planning is a cognitively demanding task (König et al., 2021) and rubrics have been shown to decrease the subjective cognitive load in self-assessment processes (Krebs et al., 2022), a scoring rubric on lesson planning is assumed to help prospective science teachers in better responding to the challenges of lesson planning. A distinction is commonly made between holistic and analytic scoring rubrics (Sadler, 2009; Tomas et al., 2019): Holistic rubrics consist of one single scale of combined criteria evaluated simultaneously, enabling an overall evaluation in a time-consuming way. In contrast, analytic scoring rubrics list a differentiated set of criteria and make performance expectations transparent. They are therefore regarded as "the gold standard of rubrics and are a good choice when […] it is important to give students detailed feedback on their strengths and weaknesses" (Schreiber et al., 2012, p. 212). The RALP is an analytic scoring rubric displaying 24 criteria describing different progressing performance levels (Sadler, 2009; Tomas et al., 2019). In this way, the RALP is more helpful for teacher educators to give sophisticated and transparent feedback on written lesson plans than a holistic scoring rubric would.

An analytic scoring rubric is only appropriately used if all criteria match the intended purposes, for example, in terms of teachers' professional standards (Moskal & Leydens, 2000). However, as far as rubrics on science lesson planning are concerned, such objectives usually do not aim to cover lesson planning competence in general but planning scientific inquiry lessons in particular. To the best of our knowledge, only six studies published in peer-reviewed journals in science education over the last two decades use analytic scoring rubrics to analyze written lesson plans (Table 1).

In contrast to the RALP, Goldston et al. (2013) and von Kotzebue, (2022) aim at quantitative analyses by calculating values. Thus, their rubrics consist of levels without qualitative performance level descriptions that would make lesson plan quality transparent. Moreover, they explicitly focus on scientific inquiry or technology-enhanced biology teaching. Kademian and Davis' (2018) rubric assesses the extent to which preservice teachers plan to engage their students in science practices. Thus, their rubric is more generally applicable but still limited to engagement as a specific learning outcome. Only Jacobs et al. (2008) provide a generally helpful rubric beyond scientific inquiry or technological pedagogical content knowledge. Even though the authors provide a couple of arguments for validity, their rubric is only based on the authors' institutes' review protocol without explicitly using an underlying theory. Such a theoretical foundation might be necessary, though, to provide a set of criteria that does objectively cover an appropriate range of aspects that are assumed to be important in written lesson plans (e.g., the CODE-PLAN model; König et al., 2021). Neither Forbes and Davis (2010) nor Enugu and Hokayem (2017) nor Kademian and Davis (2018) nor von Kotzebue, (2022) refer to validity issues but exclusively to interrater reliability.

TABLE 1 Overview of science education studies using an analytic scoring rubric in research on lesson planning.

	References					
Description	**Jacobs et al. (2008)**	**Forbes and Davis (2010)**	**Goldston et al. (2013)**	**Enugu and Hokayem (2017)**	**Kademian and Davis (2018)**	**von Kotzebue (2022)**
Applicability	Generic	Scientific inquiry	scientific inquiry	Scientific inquiry	Engaging students in scientific practices	TPACK
Structure	20 criteria 3 weighted plds	5 criteria 4 plds	21 criteria 5 levels without plds	9 criteria 6 levels without plds	10 criteria 3 plds	5 criteria on instructional quality 3–4 levels without plds; 1 criterion on the implementation of digital technologies 4 levels without plds
Reliability	inter (%)	inter (%)	inter (%)	inter (%)	inter (%)	inter (κ)
Validity	Correlation between rubric scores and scores from a lesson observation protocol and a questionnaire, respectively	–	Expert discussions, factor analysis	–	–	–
Qualitative insights into examplary lesson plans	–	–	–	A good, a mediocre, and a poor example for two aspects of scientific inquiry each	Strengths and weaknesses in six participants' lesson plans	–

Abbreviations: inter (κ), interrater agreement in terms of Cohen's κ; inter (%), interrater agreement in terms of percentage; plds, performance level descriptions; TPACK, technological pedagogical content knowledge; –, indicates that this aspect is not considered in the referenced study.

Beyond science education, it seems to be common to focus on interrater reliability and at least validity evidence based on test content (e.g., expert discussions) in the development of scoring rubrics on lesson planning (see Backfisch et al., 2020; Ndihokubwayo et al., 2022; Ruys et al., 2012). For instance, Ndihokubwayo et al. (2022) draw the following conclusion regarding the validity of their interdisciplinary *lesson plan analysis protocol* (LPAP):

> The study's design was motivated by the gap identified in the lack of tools to analyze pedagogical documents such as lesson plan [sic!] that reflect on the competency-based pedagogy. The draft protocol was developed first, together with its training manual [...]. The produced initial protocol was sent to different experts with considerable experience in the competence-based curriculum for validity purposes; this was done and yielded an improved version of the LPAP. After this stage, the reliability check process started; through this process, a very good LPAP was produced. The LPAP is a valid and reliable tool for teachers and educational evaluators. (p. 6)

In this paper, we aim to build a validity argument for the RALP, and thereby, we illustrate that testing an instrument for validity is a challenging and complex endeavor (Kane, 2013); it is even understood as "the quest for the Holy Grail" (Scherer, 2017). Testing for validity needs to cover multiple aspects of validity (AERA, APA, & NCME, 2014) to build a sound and convincing validity argument and, thus, ensure that the scores can be interpreted validly.

3.3 | Toward a validity argument for the RALP

To ensure the quality of a scoring rubric, issues regarding objectivity, reliability, and validity should be considered in the rubric development process (Brookhart, 2018; Jonsson & Svingby, 2007; Moskal, 2000; Pellegrino et al., 2016). Mainly validity issues are often insufficiently addressed (Table 1; Reddy & Andrade, 2010). Validity "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). It is a crucial prerequisite for the interpretation of empirical findings, yet it "is not a property of the test[,] [r]ather, it is a property of the proposed interpretations and uses of the test scores" (Kane, 2013, p. 3). Kane (1992) outlined the "argument-based approach to validity": According to him, validation is characterized as a process of gaining empirical evidence and building an argument for the validity of the interpretations of the test scores. Three steps will be carried out in this study (Figure 2):

3.3.1 | Step I: State the validity argument

First, it is necessary to make the validity argument or "interpretation/use argument"[2] (Kane, 2013, p. 2) transparent (Figure 2) and clearly state "why, how, by whom and in what contexts [the scoring rubric is intended to be] used" (Turley & Gallagher, 2008, p. 87): In the present study, we aim to ensure valid interpretations of the RALP scores. The RALP is intended to summatively assess the quality of preservice science teachers' (PSTs') and trainee science teachers' (TSTs') lesson plans (Supporting Information: Appendix; Großmann & Krüger, 2022a) within the framework of PCK (Carlson et al., 2019; Park & Oliver, 2008). More specifically, we intend for PSTs' teacher educators in the first and in the second phase of German teacher training to use the rubric as a framework for providing profound and differentiated feedback on their prospective science teachers' lesson plans. Consequently, PSTs' and TSTs' scores are interpreted as indicators of the lesson plan quality. For the sake of terminological clarity: In this study, we target prospective science teachers' ability to write proper lesson plans, which comprises both general lesson planning competence that is equally necessary for everyday lesson planning and the ability to justify their decisions.

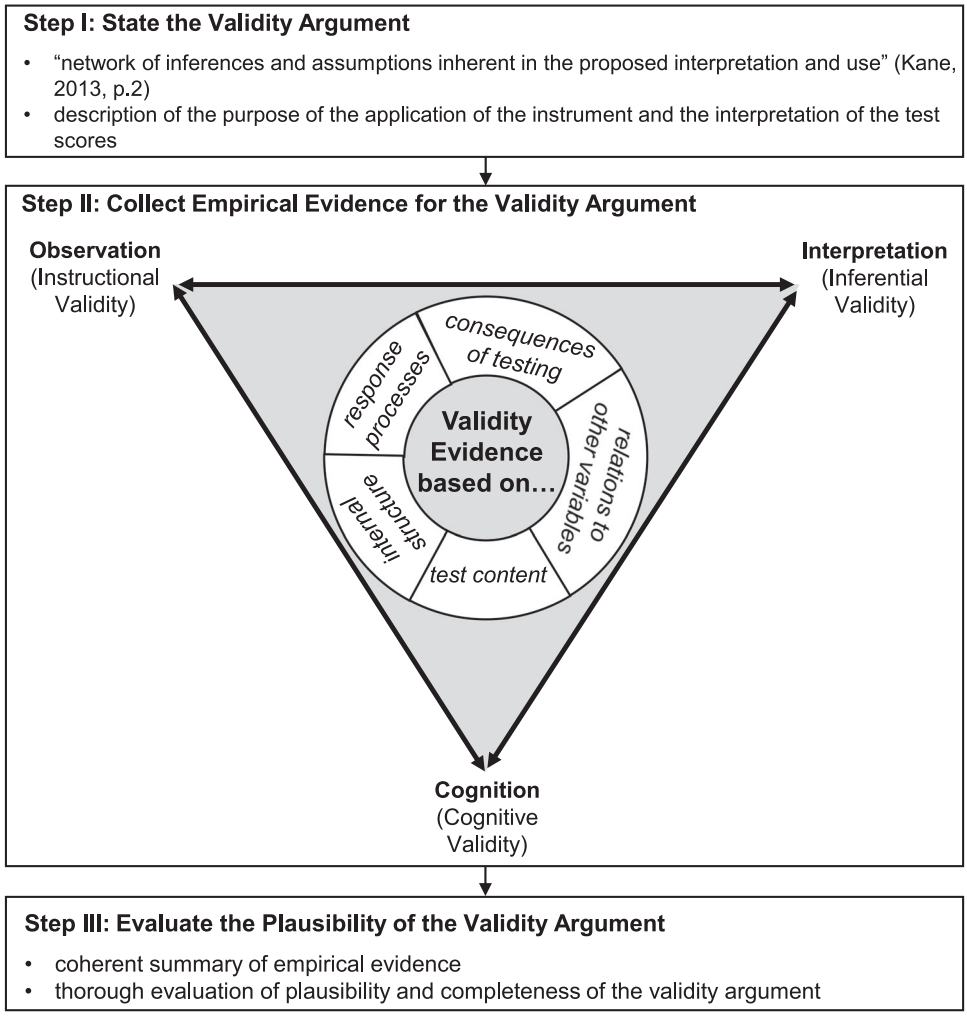

FIGURE 2 Building a validity argument in three steps (based on Kane, 1992, 2013). *Assessment triangle* was adapted from Pellegrino et al. (2016) and the allocation of the five sources of validity evidence from AERA, APA, & NCME (2014).

3.3.2 | Step II: Collect empirical evidence for the validity argument

Second, collecting empirical evidence that supports the validity argument is necessary (AERA, APA, & NCME, 2014; Kane, 2013). To build a sound, coherent, and comprehensive validity argument and cover a reasonable range of empirical evidence, we merged two frameworks of different grain-size that are concerned with validity issues: (1) Pellegrino et al.'s (2016) coarse-grained *Assessment Triangle* and (2) AERA, APA, & NCME, (2014) fine-grained "Standards for educational and psychological testing" (Figure 2).

(1) Even though Pellegrino et al.'s (2016) framework explicitly refer to assessments close to classroom instruction, it appears to be eligible also in the context of lesson planning as a part of instruction in higher education (Drost & Levine, 2015; Karlström & Hamza, 2021). When applied in this context, the framework offers the advantage of strengthening the importance of an assessment instrument to (i) reflect the

knowledge domains to be assessed (*Cognition*; Figure 2). If the interpretation of the RALP scores reflects solely the ability to write a good lesson plan (and no other knowledge domains or competencies), the RALP would be characterized by a high degree of *cognitive validity*; (ii) be aligned to the professional standards in the curriculum and the teaching and learning before the assessment (*Observation*; Figure 2). If the RALP performance level descriptions match the intended learning outcomes regarding lesson planning in teacher training and cover the aspects of lesson planning that are part of instruction in teacher training, the RALP would be characterized by a high degree of *instructional validity*; (iii) yield reasonable conclusions about the observations made concerning reliability, appropriateness, and generalizability (*Interpretation*; Figure 2). If evidence can be collected for the RALP scores to provide reliable inferences about prospective teachers' competence to write a lesson plan employing statistical analyses, the RALP would be characterized by a high degree of *inferential validity*.

(2) Within this *Assessment Triangle* (Pellegrino et al., 2016), different specific sources of valid evidence need to be taken into consideration (AERA, APA, & NCME, 2014):
   - *Validity evidence based on test content* is needed to ensure that the RALP criteria are relevant and match teacher educators' expectations of good lesson plans (e.g., by expert ratings).
   - *Validity evidence based on response processes* is needed to ensure the target group can use the RALP (e.g., by documenting teacher educators' comprehension difficulties).
   - *Validity evidence based on internal structure* is needed to ensure that the RALP criteria show sufficient psychometric properties regarding the underlying conceptual frameworks of PCK and lesson planning (e.g., dimensionality, item discrimination).
   - *Validity evidence based on relations to other variables* is needed to ensure that the RALP scores correlate with scores from different methodological approaches that measure PCK or lesson planning competence. For instance, the experience might be a variable to test for, so comparing different samples that vary in their planning experience is necessary.
   - *Validity evidence based on the consequences of testing* is needed to ensure the interpretations of the RALP scores are appropriate for their intended use. For instance, high RALP scores are expected to reflect high lesson plan quality, and vice versa (i.e., "predictive validity" in outdated terminology of educational measurement), and PSTs/TSTs are expected to be able to improve their lesson plans after having received detailed formative feedback by their teacher educators. Moreover, high scores might be expected to correlate with the quality of teaching the lesson.

3.3.3 | Step III: Evaluate the plausibility of the validity argument

Finally, it is necessary to summarize the various pieces of empirical evidence and evaluate to what extent it is clear, coherent, and plausible and supports the validity argument. Kane (2013) points out that

> the evaluation of evidence in the validity argument is not symmetric. To make a positive case for the proposed interpretations and uses of scores, the [validity argument] needs to provide adequate backing for all of the inferences […] and to rule out challenges based on plausible alternative interpretations. However, a refutation of any core warrant can be decisive in undermining [a validity argument]. (p. 16)

The purpose of this study is to build a validity argument for the interpretations of the RALP as proposed by Kane (2013), whose argument-based approach demands researchers to lay out the validity argument and collect empirical evidence for it. Applied to the RALP, Table 2 shows multiple validity arguments for all three vertices of the

TABLE 2 Validity arguments and related sources of validity evidence for the RALP.

Component of validity	Validity argument	Validity evidence (*based on…*)
1. Cognition (cognitive validity)	The RALP criteria reflect the expected cognitive construct (i.e., PCK needed for lesson planning) if … – the criteria unpack relevant aspects of written lesson plans in science education. – the PCK components and their interconnections are aligned with relevant aspects of lesson planning.	1a. The RALP criteria build upon theoretical and empirical research on lesson planning in science education (*test content*). 1b. Science teacher educators in the first phase of teacher education judge the theoretical and empirical foundation of the RALP criteria (*test content*).
2. Observation (instructional validity)	The observed RALP scores reflect the professional standards with regard to lesson planning if … – Teacher educators consider the criteria meaningful and relevant for writing lesson plans in teacher training. The RALP criteria are suitable components of an analytic scoring rubric if … – the criteria interconnect different PCK components. – the target group can meet the criteria. – the criteria help discriminate between high- and low-performing PSTs and TSTs, respectively.	2a. Science teacher educators in the second phase of teacher education judge that the RALP criteria are relevant in teacher training, that is, whether the rubric is fully aligned with the performance expectations at the end of teacher training (*test content*). 2b. The manifest RALP criteria cannot be reduced to a small set of meaningful latent variables employing a dimensionality analysis (*internal structure*). 2c. The item difficulty indicates that the RALP criteria are neither too difficult nor too easy to meet for PSTs and TSTs, respectively (*internal structure*). 2d. The item discrimination indicates that the PSTs'/TSTs' overall RALP score correlates to their success on the single criteria (*internal structure*).
3. Interpretation (inferential validity)	The observed RALP scores account for science teachers' ability to plan lessons sophisticatedly if… – scores predict the lesson plan quality in terms of a holistic quality assessment by teacher educators. – the scores are sensitive to differences between PSTs and TSTs.	3a. RALP scores (0–48) and the holistic quality assessment by experienced teacher educators in terms of grades (1—very good, 2—good, 3—satisfactory, 4—sufficient, 5–poor) are negatively correlated (*consequences of testing*). 3b. TSTs at the end of teacher training achieve higher RALP scores than PSTs (*relations to other variables*).

Note: Procedure adapted from Zhai et al. (2021); components of validity derived from Pellegrino et al. (2016); sources of validity evidence derived from AERA et al. (2014).

Abbreviations: PSTs, preservice science teachers; RALP, rubric to assess science lesson plans; TSTs, trainee science teachers.

Assessment Triangle (Pellegrino et al., 2016) and the corresponding sources of validity evidence (AERA, APA, & NCME, 2014). It is important to note that "each type of evidence presented [above] is not required in all settings. Rather, support is needed for each proposition that underlies a proposed test interpretation for a specified use" (AERA, APA, & NCME, 2014, p. 14). In this study, we focus on some aspects relating to the relevance of the RALP criteria and their potential to assess lesson plan quality properly in a summative sense. By now, we focus on

ensuring that the RALP scores reflect lesson plan quality. We have not yet intended to investigate how far teacher educators can apply the RALP. This might be a subsequent step in the validation process. Hence, this paper does not provide information concerning *validity evidence based on response processes*.

4 | PROBLEM STATEMENT AND RESEARCH QUESTIONS

Although lesson plans are crucial, for example, in the German teacher training system (KMK, 2019; König et al., 2020; Neumann et al., 2017), a theoretically framed scoring rubric applicable to all kinds of science lessons and tested for validity (Table 1) comprising empirically tested criteria to evaluate the quality of written lesson plans (Krepf & König, 2022) are still missing. Such a rubric would offer a twofold advantage: First, it might facilitate comparability between teacher educators' assessment of lesson plan quality in teacher education and thus reduce the impact of teacher educators' subjective preferences. Second, it would allow researchers in science education to use an established and evaluated instrument to assess lesson plan quality in qualitative and quantitative studies.

This study pursues two objectives: First, we aim to evaluate the RALP quality by building a valid argument and collecting empirical evidence (Table 2). The following research questions will be addressed:

RQ 1.1: To what extent do science teacher educators consider the RALP criteria relevant for writing lesson plans? (*Validity evidence based on test content*).

RQ 1.2: To what extent does the scoring rubric allow for reliable measures?

RQ 1.2a: To what extent is the RALP scoring procedure stable over a 2-week interval when applying the scoring rubric to PSTs' and TSTs' lesson plans? (*Intrarater reliability*).

RQ 1.2b: To what extent do a science teacher educator and a trained student assistant reach an intersubjective agreement when applying the RALP to TSTs' lesson plans? (*Interrater reliability*).

RQ 1.3: To what extent do data reflect the intended interconnectedness of PCK components? (*Validity evidence based on internal structure*).

RQ 1.4: To what extent is the selection of criteria appropriate to assess lesson plan quality? (*Validity evidence based on internal structure*).

RQ 1.5: To what extent do the RALP scores correlate with science teacher educators' holistic quality assessment? (*Validity evidence based on the consequences of testing*).

RQ 1.6: To what extent do PSTs and TSTs achieve different scores? (*Validity evidence based on relations to other variables*).

Second, after ensuring that RALP scores can be interpreted validly, we aim to provide a qualitative insight into prospective science teachers' lesson planning competence:

RQ 2: What are the most significant differences between PSTs' and TSTs' lesson plans?

5 | METHODS

5.1 | Study context

German science teacher education is structured both concurrently and consecutively (Kotthoff & Terhart, 2013): It is concurrent because, from the very beginning, prospective teachers are confronted with all relevant elements (educational studies, CK and PCK in their two subjects, school practice, planning, teaching, reflecting). It is consecutive since teacher training consists of two distinct phases (Figure 3): The first phase is located at university (*PSTs*), is focused on theory, and lasts 5 years (Bachelor: 3 years, Master: 2 years). During the Master's studies, PSTs must complete a 5-month practical semester at school, where a science teacher educator visits them to give feedback on their planning, teaching, and reflection competence (Figure 3b). At the end of the practical semester,

and after being taught how to write lesson plans, PSTs write a lesson plan for a graded paper. The second phase is the induction phase, located at schools (*in-service trainee science teachers*, *TSTs*) accompanied by state-run seminars, and lasts another 1.5 years (Kotthoff & Terhart, 2013; Neumann et al., 2017). During the second phase, science teacher educators visit the TSTs at schools five times per subject to give feedback on their planning, teaching, and reflection competence (Figure 3c). As their final examination, TSTs must plan, teach, and reflect on one lesson each in both their subjects. In our case, their performance was only evaluated based on a lesson plan discussion, as the Covid-19 pandemic made it impossible to assess their teaching. An expert committee consisting of up to five science teacher educators working in the second phase of teacher education and the school principal carried out this evaluation. According to the standard certification requirements, prospective science teachers are intended to be able to plan lessons under consideration of different learning dispositions among students and teach effectively. More specifically, they are intended to acknowledge students' diversity, plan lessons accordingly, interconnect their CK, PK, and PCK to choose appropriate topics, media, and activities, and assess the quality of their teaching practice (Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK], 2019). If passed, TSTs receive a certificate, which qualifies them to work as full-time in-service teachers in Germany. Teacher education programs vary across the 16 federal states of Germany (Kotthoff & Terhart, 2013); the structure described here only holds for Berlin, the federal state where we conducted this study.

5.2 | Sample and data collection

For this study, we analyzed lesson plans ($N = 100$) written by prospective biology teachers (PSTs, TSTs). This sample comprises lesson plans written by PSTs ($n = 36$) who attended the two authors' courses at university. Lesson plans ($n = 64$) written by TSTs due to their final examinations ("Second State Examination"; Kotthoff & Terhart, 2013) of

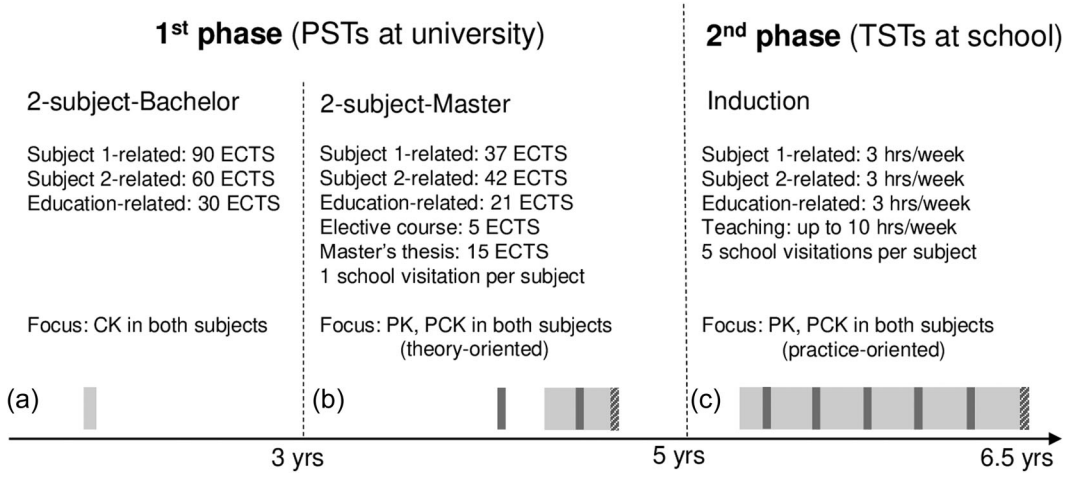

FIGURE 3 Overview of the structure of teacher training in Germany. Light gray boxes indicate practical parts of teacher training, dark gray boxes indicate instances when PSTs/TSTs have to write a lesson plan, and striped dark gray boxes indicate samples analyzed in this study: (a) 4-week orientation internship in the first year of teacher training (no lesson plan); (b) 5-month practical semester in the fourth year of teacher training (three lesson plans); (c) 1.5-year induction phase at the end of teacher training (five lesson plans and another lesson plan in the course of the final examination). The authors of this paper work in the first phase at university and thus do not have any influence on the second phase. One ECTS credit equals 30 h of workload. CK, content knowledge; ECTS, European Credit Transfer and Accumulation System; hrs, hours; PCK, pedagogical content knowledge; PK, pedagogical knowledge; PSTs, preservice science teachers; TSTs, trainee science teachers; yrs, years.

German teacher training in the years 2019–2021 (Figure 3c, striped dark gray box) were provided anonymously by the Ministry of Education. Due to that, we could not collect demographic information about the participants or the schools they teach at. Since the RALP (Supporting Information: Appendix A) is intended to display performance level descriptions up to the highest level that can be expected for each criterion, this sample can be regarded as suitable: Lesson planning competence develops in the course of teacher training (Backfisch et al., 2020; König et al., 2021, 2022; Mutton et al., 2011; Vogelsang et al., 2022; Westerman, 1991). Moreover, the focus of teacher education shifts from acquiring CK during Bachelor studies (Figure 3a) to PK and PCK during Master studies (Figure 3b) and particularly during the practice-oriented induction phase (Figure 3c; Kotthoff & Terhart, 2013). Consequently, we do not assume Bachelor's or Master's students to be able to achieve the highest levels. Moreover, the authors did not influence the development of these lesson plans since the state, and not teacher educators at the university, run the induction phase. Therefore, the TSTs followed the official guidelines of the school administration (SenBJF, 2017), which is not associated with the authors' affiliations. Hence, data collection is characterized by a high degree of ecological validity, "the extent to which research findings would generalize to settings typical of everyday life" (Wegener & Blankenship, 2007, p. 275). For the sake of clarity: The term "lesson plan" used in this study does explicitly not refer to short, 1–2 pages long sketches of lessons (e.g., Morine-Dershimer, 2011). Instead, a regular lesson plan in German teacher training covers approximately 10 pages (plus appendices).

To answer RQ 1.5, RQ 1.6, and RQ 2, a comparison to another sample is required to investigate whether the RALP is sensitive to the assumed differences in perspective science teachers' PCK and lesson planning competence (Figure 3). Therefore, lesson plans ($n = 36$) were collected from PSTs' graded papers written after their practical semester in 2019–2021 (Figure 3b, striped dark gray box). For the data analysis, these lesson plans were anonymized. Both PSTs' and TSTs' lesson plans were written for graded exams. They prepared their lesson plans with no time constraints out of school and out of their seminars. Consequently, the lesson plans were not written under controlled conditions. We have no information on how long the PSTs and TSTs worked on their lesson plans, which resources they used, and whether they asked colleagues for feedback.

5.3 | Data analysis

Using Creswell and Plano Clark's (2018, pp. 77–84) denotation, our approach could be described as a QUAN → qual explanatory sequential design that aims to understand better the quantitative data used for evaluating validity. Consequently, we first analyzed mainly quantitative data collected by application of the RALP (RQ 1) and second shed light on selected relevant aspects in qualitative analysis (RQ 2). In detail, the data were analyzed as described below.

To ensure that the criteria are relevant for writing lesson plans (RQ 1.1; *Validity evidence based on test content*; AERA, APA, & NCME, 2014), we led expert discussions with the two authors' colleagues and afterward individual discussions with two professors of biology education, both working in different federal states of Germany. Since all federal states are individually responsible for their education policy in Germany, we thus aimed to increase the generalizability of the RALP scores. To ensure that teacher educators consider the criteria relevant, all 25 teacher educators working in the second phase in the same federal state as the authors of this paper (Berlin) were asked via e-mail to participate in a survey. 10 teacher educators agreed to participate. They were asked to rate the relevance of an initial set of 29 criteria on a 4-point Likert scale (1—*irrelevant*, 2—*barely relevant*, 3—*rather relevant*, 4—*very relevant*). In contrast to the teacher educators working in the first phase at university, the teacher educators working in the second phase are in-service teachers who run seminars for TSTs and form part of the examination board for the final examination of teacher training (Figure 3). For the rating, we only presented the highest performance level description of each criterion, as this is what TSTs are expected to achieve at the end of their teacher training.

As a prerequisite for further validity analysis, we needed to ensure that the scoring procedure was objective and reliable. To address RQ 1.2a and RQ 1.2b, we calculated weighted Cohen's κ as a measure of rater agreement (Landis & Koch, 1977), taking into account the degree of disagreement between two raters (Gwet, 2014). This paper will address two kinds of reliability: (a) For intrarater reliability (RQ 1.2a), the first author applied the RALP to 18/100 randomly selected lesson plans twice in a 2-week interval. Weighted Cohen's κ was calculated. The first author resolved disagreements between his two codings and reached a final version. (b) The interrater reliability (RQ 1.2b) of two different raters (the first author and a trained student assistant) was calculated as well by using weighted Cohen's κ. Again, disagreements in codings were discussed and resolved.

Afterward, we evaluated different sources of validity evidence (AERA, APA, & NCME, 2014) of the RALP: First, to investigate the empirical structure of the RALP (RQ 1.3; *validity evidence based on internal structure*; AERA, APA, & NCME, 2014), we performed an analysis within the framework of item response theory (IRT; Bond & Fox, 2001) using the software ACER Conquest (version 1.0.0.1; Wu et al., 2007). As an extension of the simple logistic model within IRT, we applied the partial credit model (Masters, 1982) to calculate item difficulties and personal abilities because the RALP criteria are polytomous (i.e., levels 0–2). As described above, the 24 criteria of the RALP mainly address interconnections between four of the five PCK components (Park & Oliver, 2008). As we differentiated between criteria that focus on the intended learning outcomes (O1–O4) and those that focus on the biological subject matter (C1–C4) within KSC (see Supporting Information: Appendix), we divided this component into two subcomponents for the statistical analysis (i.e., KSC—focus on intended learning outcomes, KSC—focus on biological content). Thus, we differentiated five components and neglected OTS. As we intend the criteria to reflect the interconnectedness, we would not assume to find five latent dimensions in the data but rather a one-dimensional, global latent dimension (i.e., the ability to write good lesson plans). Thus, data would reflect the intended interconnectedness of PCK components. In addition, two further models were specified and compared: a two-dimensional model separating all criteria with explicit reference to the intended learning outcome and those without, and a four-dimensional model that considers the relative closeness of criteria to each other, as shown in Figure 1. To decide which model fits best, we calculated the following values based on the analysis of $N = 100$ lesson plans. Thereby, we followed an established procedure in science education research (e.g., Großschedl et al., 2014): The factor of final deviance indicates to what extent the collected data fit the underlying assumptions of each model. To test which model fits best, we calculated two descriptive estimates, namely Akaike (1981) information criterion (AIC) and the Bayesian information criterion (BIC; Schwarz, 1978). Considering the parsimony of the models, it is assumed for both AIC and BIC that the lower the value, the better the data fit the model (Wilson et al., 2008). To decide which model fits best to the data, we performed χ^2 tests. For that purpose, the difference between the deviances and the degrees of freedom as differences in the number of parameters were calculated, finally leading to an estimation of whether the two models are significantly different from each other (Bentler, 1990). Each model's reliabilities for each dimension regarding the expected a posteriori/plausible values will be reported.

Second, to gain further insight into the psychometric appropriateness of the criteria (RQ 1.4; *validity evidence based on internal structure*; AERA, APA, & NCME, 2014), the item discrimination index (a_i) indicates how far the criteria can differentiate between persons with different abilities in terms of the correlation between the persons' scores on a particular criterion and their overall score. As a rule of thumb, items exceeding the threshold of 0.30 can be regarded as sufficiently discriminatory (Vaus, 2014). In addition, we will provide a Wright map (Boone et al., 2014) illustrating how far the PSTs' and TSTs' abilities match the difficulty of the criteria. It would be desirable if the average person's abilities were close to the average item difficulty.

Third, we aimed to investigate whether a high score reflects a high overall lesson plan quality (*validity evidence based on the consequences of testing*; AERA, APA, & NCME, 2014), that is, predictive validity (RQ 1.5). For that purpose, we calculated the Pearson correlation between the teacher educators' holistic quality assessments for the 36 PSTs' lesson plans and the teacher educator's holistic quality assessments for the 64 TSTs' lesson plans on the

one hand and the RALP scores on the other. The validity argument mentioned above (Table 2) refers to the purpose of the RALP as an instrument for teacher educators to assess prospective science teachers' lesson plan quality. Consequently, the RALP needs to match teacher educators' perspectives, and its scores need to reflect teacher educators' holistic quality assessment. As established in Germany, the following range of grades was used to assess lesson plan quality holistically: 1.0/1.3—very good; 1.7/2.0/2.3—good; 2.7/3.0/3.3—satisfactory; 3.7/4.0—sufficient; 4.3/4.7/5.0—poor. As mentioned above, the authors of this paperwork in the first phase of teacher education at the university, while the state runs the second phase (Figure 3). Consequently, the holistic quality assessments of PSTs' and TSTs' lesson plans differed slightly: PSTs' lesson plans ($n = 36$) were assessed by six teacher educators working in the first phase at a German university (three professors, two postdoctoral researchers, one PhD student). One professor and the PhD student are the authors of this paper. To ensure an objective grading procedure, the six teacher educators used the same grading scheme with various criteria that PSTs were expected to consider in their lesson plans (Table 3). TSTs' lesson plans ($n = 64$) were assessed by a retired biology teacher who served as a teacher educator in the second phase for 21 years (Figure 3). The teacher educator was instructed to grade the 64 lesson plans based on the holistic criteria he applied (Table 3) when he was still in service and part of official examination committees. In contrast to the PSTs, TSTs were not explicitly expected to consider students' conceptions. Still, they were instructed to provide a tabular unit and clarify what competencies they wanted to foster during this unit.

Two steps were taken to test the teacher educator's objectivity and trustworthiness. (1) We calculated the Pearson correlation between the teacher educator's grades and the original grades we obtained for no more than 17 of those 64 lesson plans provided by the Ministry of Education. (2) We hired a second teacher educator working in the second phase who graded a subsample of lesson plans as already described in Großmann and Krüger (2022c) and calculated the Pearson correlation coefficient again. Moreover, the teacher educator was asked to elaborate on all 64 lesson plans and give insights into the reasons for the grades he decided to give in a series of interviews. Both teacher educators were unfamiliar with the authors' affiliations and were unfamiliar with the RALP. Thus, we ensured that the holistic quality assessment could serve as a valid external criterion to evaluate the quality of the RALP.

PSTs have to meet the criteria established by their teacher educators at the university; TSTs have to meet the official guidelines for the second phase of teacher training and the final examination of teacher education (SenBJF, 2017).

TABLE 3 Requirements of written lesson plans.

Requirements	PSTs	TSTs
Analysis of biological subject matter	x	x
Intended learning outcome	x	x
Unit contextualization		x
Students' level of competence (internal differentiation)	x	x
Students' conceptions of the subject matter	x	
Description and justification of planning decisions (e.g., tasks, activities)	x	x
Tabular schedule of the lesson	x	x
Appendices (e.g., worksheets)	x	x

Note: The requirements differ slightly between PSTs' lesson plans ($n = 36$) and TSTs' lesson plans ($n = 64$). PSTs have to meet the criteria established by their teacher educators at the university; TSTs have to meet the official guidelines for the second phase of teacher training and the final examination of teacher education (SenBJF, 2017).

Abbreviations: PSTs, preservice science teachers; TSTs, trainee science teachers.

Since holistic grading might always be subjective to a certain extent, we aimed to decrease the significance of these grades by dividing both subsamples into quartiles based on the holistic quality assessment: The top 25% ($n_{PSTs} = 9$; $n_{TSTs} = 16$), the middle 50% ($n_{PSTs} = 18$; $n_{TSTs} = 32$), and the bottom 25% ($n_{PSTs} = 9$; $n_{TSTs} = 16$) were compared concerning the RALP scores they achieved. Thus, we focus on the ranks within the sample rather than on the teacher educator's exact grades. Mann–Whitney *U* tests were calculated to determine whether these quartiles' differences are significant.

Fourth, to investigate the instructional sensitivity of the RALP (RQ 1.6; *validity evidence based on relations to other variables*; AERA, APA, & NCME, 2014), the Mann–Whitney *U* test was applied to compare whether the two subsamples of PSTs ($n = 36$) and TSTs ($n = 64$) differ significantly regarding their RALP scores.

After ensuring that the RALP scores allow for valid interpretations of scores, we intend to provide qualitative insights into prospective science teachers' lesson planning competence. More specifically, based on the assumption that lesson planning competence develops during teacher education (Backfisch et al., 2020; König et al., 2021, 2022; Mutton et al., 2011; Vogelsang et al., 2022; Westerman, 1991), we aim to illustrate in detail how PSTs and TSTs approach significant challenges in the process of lesson planning. First, we compared the PSTs' scores with the TSTs' scores, which were generally higher, and identified the three criteria showing the largest effect sizes. Second, we applied the *typical case sampling strategy* (Patton, 1990, pp. 173–174): We purposefully selected illustrative extracts from TPSTs' and PTSTs' lesson plans for each of the three criteria. With these cases, we were especially interested in differences in how PSTs and TSTs tend to plan a lesson and justify their decisions.

6 | RESULTS

6.1 | Relevance of criteria (RQ 1.1; *test content*)

The teacher educators' feedback provides evidence that the RALP criteria are well-aligned with professional standards and cover a range of relevant lesson-planning aspects. Since terminology is sometimes equivocal in German educational literature, some expressions were changed to be more generally comprehensible. In particular, feedback helped us describe sophisticated performances at the highest level.

Most criteria were considered rather relevant by the 10 teacher educators (Table 4). Two criteria were considered irrelevant (C2: use of references to educational literature, $M = 1.45$; S3: consideration of students' conceptions, $M = 1.55$). C2 was refined so that the focus now lies on the topic's relevance rather than how prospective teachers justify their choice; it no longer forms a single criterion but was added to the highest performance level description for some criteria. S3 was not deleted due to its relevance in research on science teacher education. More importantly, several teacher educators stressed that students' conceptions might be necessary for some lessons but not all. We do not consider this a convincing argument and thus decided to maintain S3 in the RALP. Furthermore, teacher educators suggested adding the selection of technical terms (C4) as an essential aspect of lesson planning, provided ideas for more precise formulations, and suggested further merges of criteria, which led to a reduction of 29 initial criteria to 24. These remaining criteria are shown in Supporting Information: Appendix A and were used for further data analysis.

6.2 | Stability and objectivity (RQ 1.2a, RQ 1.2b; *intra- and interrater reliability*)

On average, weighted Cohen's κ indicated "almost perfect" intra- ($M = 0.93$, range = 0.77–1.0) and interrater agreement ($M = 0.88$, range = 0.65–1.0; Landis & Koch, 1977, p. 165), indicating that the scoring procedure is both stable and intersubjectively comprehensible. In addition, a criteria-wise analysis of the reliability coefficients (Table 4) led to the identification of single criteria that were difficult to apply for the first coder (e.g., C1, I3, I7) or

TABLE 4 Teacher educators' relevance assessment (mean and standard deviation) on a 4-point Likert scale, intrarater reliability, interrater reliability (weighted Cohen's κ), and absolute frequency of the final codings, sorted by criterion.

RALP criteria		M_{TE}	SD_{TE}	κ_{intra}	κ_{inter}
O1	Intended learning outcome in the unit	2.18	1.00	0.89	0.89
O2	Progression throughout the unit	1.91	0.89	0.91	0.82
O3	Intended learning outcome in the lesson	2.91	0.30	0.89	0.58
O4	Indicators as evidence of the desired learning	2.64	0.49	0.91	0.72
C1	Analysis of biological content	2.00	0.60	0.70	0.76
C2	Choice of topic	1.45	0.92	0.91	0.82
C3	Educational reconstruction	1.91	1.00	0.92	1.00
C4	Selection of relevant technical terms	n/a	n/a	0.91	0.69
S1	Analysis of the level of competence	2.27	0.50	1.00	0.83
S2	Progression of competence development	2.82	0.30	1.00	0.67
S3	Students' conceptions	1.55	0.80	1.00	1.00
S4	Learning difficulties	2.64	0.46	1.00	0.88
S5	Methodical skills	2.64	0.64	1.00	0.83
I1	Structure of the development of competencies	2.70	0.46	0.88	0.57
I2	Lesson structure	n/a	n/a	1.00	0.80
I3	Suitability of the methods	2.90	0.30	0.20	0.79
I4	Suitability of tasks	2.50	0.50	1.00	0.81
I5	Suitability of materials	2.70	0.46	0.87	0.76
I6	Horizon of expectations	2.50	0.67	0.83	1.00
I7	Adaptive teaching	2.90	0.30	0.70	0.78
I8	Anticipated difficulties and alternatives	2.30	0.46	1.00	1.00
I9	Transparency of the learning process	2.80	0.40	1.00	1.00
A1	Transparency of performance expectations	2.56	0.68	0.91	0.71
A2	Products of students' learning	2.70	0.64	0.77	0.89

Note: C4 and I2 were added in response to the teacher educators' feedback, so no Likert scale results are reported.
Abbreviations: n/a, not applicable; RALP, rubric to assess science lesson plans.

the second coder (e.g., O3, C4, I1). The formulations of these criteria were then further checked for imprecise or confusing expressions and clarified.

6.3 | Dimensionality (RQ 1.3; *internal structure*)

The decisive information-based criteria AIC and BIC are lower for the one-dimensional model than for the remaining three models (Table 5). χ^2 tests show that the one-dimensional model fits the data significantly better than the two-dimensional model (χ^2 [2] = 7.72, $p < 0.05$). In contrast, there is no significant difference to the

TABLE 5 Fit statistics for the four PCM models.

Model	Parameters	Deviance	AIC	BIC	EAP/PV
One dimensional	25	4859.34	4909.34	4974.46	0.50
Two dimensional	27	4867.05	4921.05	4991.39	0.48/0.53
Four dimensional	34	4860.59	4928.59	5017.16	0.57/0.45/0.58/0.32
Five dimensional	39	4869.22	4947.22	5048.82	0.59/0.49/0.45/0.62/0.36

Abbreviations: AIC, Akaike (1981) information criterion; BIC, Bayesian information criterion (Schwarz, 1978); EAP/PV, expected a posteriori/plausible values; PCM, partial credit model.

four-dimensional (χ^2 [9] = 1.25, p = 0.99), and the five-dimensional model (χ^2 [14] = 9.88, p = 0.77). The psychometric analysis supports the assumption that a one-dimensional model fits the data best.

6.4 | Appropriateness of criteria (RQ 1.4; *internal structure*)

Overall, absolute codings for each criterion show that all criteria are achievable in both subsamples (Table 6). With the exceptions of unit contextualization (O1, O2) in PSTs' lesson plans and considering students' conceptions in TSTs' lesson plans (S3), levels 1 and 2 were achieved by many prospective science teachers in both subsamples.

All but four criteria exceed the threshold of 0.30 and can thus be regarded as sufficiently discriminatory (Vaus, 2014). The remaining criteria (O4, S3, S5, and I4) barely help discriminate PSTs' and TSTs' performances.

The Wright map (Figure 4) gives further insights into the relationship between the difficulty of the proposed criteria and the target group's abilities. Regarding the difficulty of the 24 criteria (right side), three observations are noteworthy: First, most criteria lie around the middle of the distribution. Hence, they can be assumed to be moderately difficult. Second, there is a gradient of difficulty. Hence, some criteria are rather difficult (e.g., I9, I4, O2), and some are relatively easy (e.g., I3, I2, O3). Third, no persons in this sample have a 50% solution probability for the two criteria, students' conceptions (S3) and horizon of expectations (I6). Hence, we conclude they are either too difficult or too easy.

6.5 | Holistic quality assessment (RQ 1.5; *consequences of testing*)

The teacher educators in the first phase and the external teacher educator in the second phase provided a holistic quality assessment in the form of grades for all lesson plans (N = 100). A second teacher educator graded a subsample and provided a detailed qualitative review to ensure intersubjective comprehensibility and reduce the impact of the external teacher educator's subjective preferences and biases in grading. Their grades only differed slightly. We learned from their reviews that they appreciated and criticized the same aspects. Different grades resulted from the different weights they gave to certain aspects. Moreover, another subsample of 17 lesson plans written by TSTs and graded by the official examination committee correlates moderately with the teacher educator's grades (r = 0.43, p = 0.08). Based on the p value, the null hypothesis (i.e., there is no correlation between the grades) cannot be rejected. It can be assumed that calculating the correlation for more lesson plans might result in a significant relationship. However, we did not obtain more than 17 lesson plans from the Ministry of Education to check that. Since the two teacher educators reached intersubjective comprehensibility, we interpret this as sufficient evidence for convergent validity, which allows us to trust the teacher educator's grades in terms of objectivity.

TABLE 6 Absolute codings for PSTs ($n = 36$) and TSTs ($n = 64$) for levels 0–2 per criterion and item discrimination (a_i).

RALP criteria		PSTs			TSTs			
		0	1	2	0	1	2	a_i
O1	Intended learning outcome in the unit	33	1	2	3	30	31	0.64
O2	Progression throughout the unit	34	2	0	9	36	19	0.68
O3	Intended learning outcome in the lesson	4	15	17	2	35	27	0.33
O4	Indicators as evidence of the desired learning	5	20	11	11	35	18	0.26
C1	Analysis of biological content	3	26	7	10	46	8	0.34
C2	Choice of topic	16	14	6	16	15	33	0.45
C3	Educational reconstruction	11	22	3	19	34	11	0.36
C4	Selection of relevant technical terms	13	18	5	6	52	6	0.38
S1	Analysis of the level of competence	13	14	9	16	18	30	0.46
S2	Progression of competence development	5	26	5	4	28	32	0.57
S3	Students' conceptions	16	12	8	60	3	1	−0.14
S4	Learning difficulties	16	17	3	9	44	11	0.54
S5	Methodical skills	2	22	12	10	31	23	0.21
I1	Structure of the development of competencies	11	17	8	1	40	23	0.43
I2	Lesson structure	2	20	14	2	34	28	0.45
I3	Suitability of the methods	2	20	14	0	25	39	0.48
I4	Suitability of tasks	14	14	8	38	18	8	0.22
I5	Suitability of materials	5	25	6	0	48	16	0.44
I6	Horizon of expectations	7	11	18	5	7	52	0.52
I7	Adaptive teaching	14	16	6	3	20	41	0.48
I8	Anticipated difficulties and alternatives	11	23	2	9	43	12	0.40
I9	Transparency of the learning process	30	4	2	22	38	4	0.43
A1	Transparency of performance expectations	18	15	3	9	46	9	0.38
A2	Products of students' learning	14	18	4	2	50	12	0.33

Abbreviations: PSTs, preservice science teachers; RALP, rubric to assess science lesson plans; TSTs, trainee science teachers.

Overall, there is a strong correlation between the RALP scores and the teacher educators' holistic quality assessment for the PSTs ($r = -0.69$, $p < .001$), as well as for the teacher educator's holistic quality assessment for the TSTs ($r = -0.72$, $p < 0.001$). Both relationships are negative because low values for the holistic quality assessment indicate high quality (grades 1.0–2.3), and higher values indicate low quality (grades 4.3–5.0). Hence, a high RALP score correlates to a low value for the holistic quality assessment and vice versa. To gain further insight into the ability of the RALP to detect quality differences among lesson plans, we divided both subsamples into the 25% of highest, the 50% of mediocre, and the 25% of poorest lesson plan quality based on the teacher educators' holistic quality assessment. These quality groups were plotted against the overall score reached by application of the RALP (Figure 5).

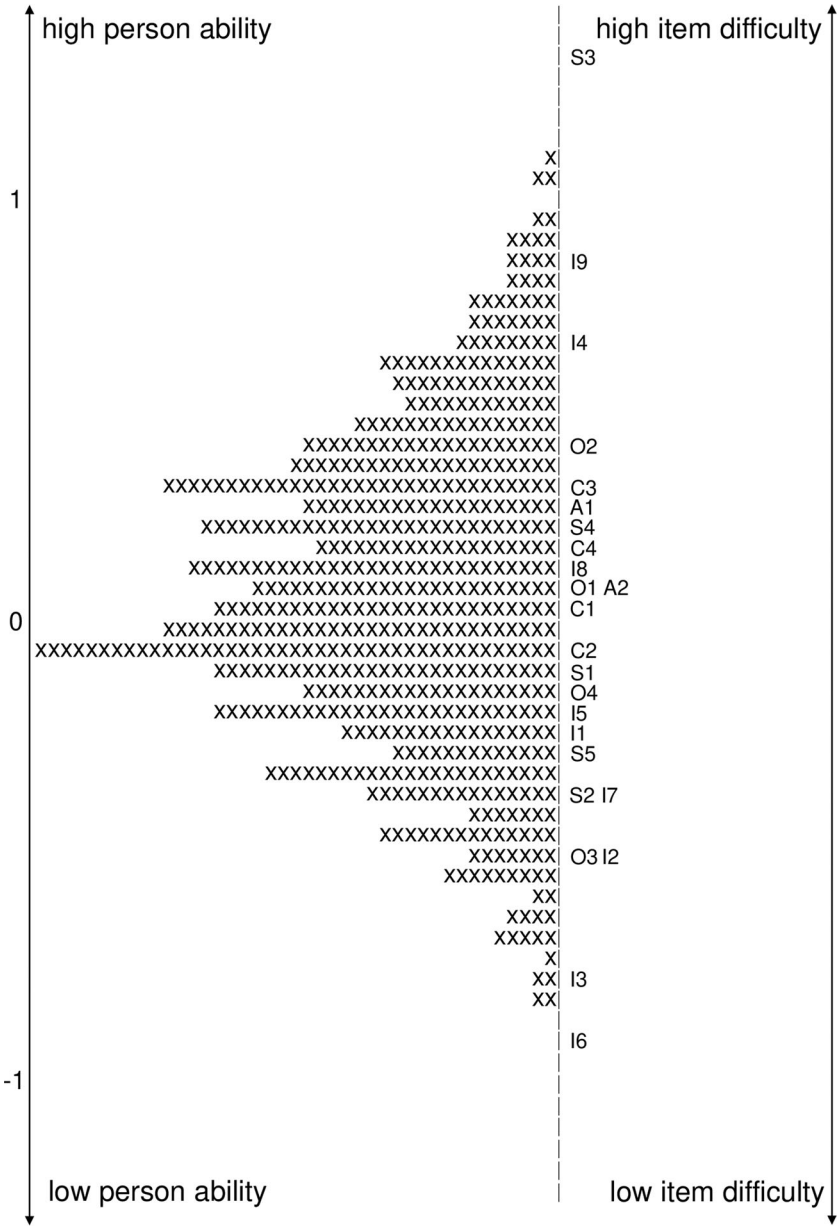

FIGURE 4 Wright map for the analysis of lesson plans ($N = 100$). On the left side, the person's abilities are shown. Each x equals 0.2 PSTs/TSTs. The closer an x is on the top, the more able the person is, and vice versa. On the right side, the item difficulties are shown. Criteria at the bottom represent those that are easier to achieve and vice versa. If an x occurs at the same position as a criterion, the person has a 50% probability of answering the item correctly. PSTs, preservice science teachers; TSTs, trainee science teachers.

Within the PST subsample, two groups of lesson plan quality differ significantly from each other: The top 25% ($Mdn = 22$) scored significantly higher than the bottom 25% ($Mdn = 15$) in the RALP ($U = 5$, $z = -3.15$, $p < 0.001$, $d = 2.28$; large effect), and the latter performed significantly worse than the middle 50% ($Mdn = 21$; $U = 6.5$, $z = -3.851$, $p < 0.001$, $d = 2.18$; large effect).

Within the TST subsample, all three groups of lesson plan quality differ significantly in their RALP scores: The top 25% (*Mdn* = 33) scored significantly higher than the middle 50% (*Mdn* = 28) in the RALP ($U = 89.5$, $z = -3.65$, $p < 0.001$, $d = 1.24$; large effect), and the middle 50%, in turn, scored significantly higher than the bottom 25% (*Mdn* = 20.5) in the RALP ($U = 66.5$, $z = -4.2$, $p < 0.001$, $d = 1.49$; large effect). As a logical consequence, the difference between the top 25% and the bottom 25% is highly significant ($U = 2$, $z = -4.76$, $p < 0.001$, $d = 3.09$; large effect).

6.6 | Instructional sensitivity (RQ 1.6; *relations to other variables*)

Comparing both subsamples of PSTs ($n = 36$) and TSTs ($n = 64$), we find that TSTs (*Mdn* = 28) score significantly higher in the RALP than the PSTs (*Mdn* = 20; $U = 421$, $z = -5.26$, $p < 0.001$, $d = 1.23$; large effect). Descriptive statistics are shown in Table 7.

We find that scores range between 11 and 39, indicating that even low-performing participants reach 25% of the maximum score. In contrast, the highest-performing participant only reaches 81.25% of the maximum score.

6.7 | Exemplary cases (RQ 2)

Mann–Whitney *U* tests indicate that TSTs score significantly higher in 13 of the 24 criteria (O1, O2, C2, S2, S4, I1, I3, I6, I7, I8, I9, A1, and A2) than PSTs ($0.41 < d < 2.01$), and PSTs score significantly higher only in S3 than TSTs. Among those criteria, the effect sizes for O1 ($d = 2.01$), O2 ($d = 1.85$), and S3 ($d = 0.90$) can be explained with the

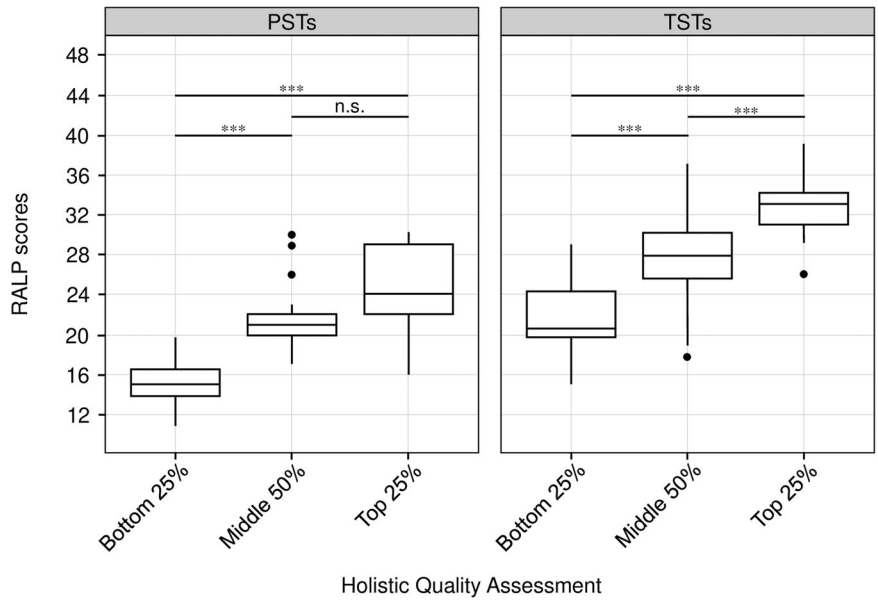

FIGURE 5 Rubric to assess science lesson plans (RALP) scores of preservice science teachers (PSTs) and trainee science teachers (TSTs) per lesson plan quality group based on the holistic quality assessment. Sample size PSTs: bottom 25% and top 25% $n = 9$; middle 50% $n = 18$; sample size TSTs: bottom 25% and top 25% $n = 16$; middle 50% $n = 32$. Possible RALP scores range between 0 and 48. Thick lines in boxes indicate medians and lower and upper box boundaries indicate 25th and 75th percentiles. Lower and upper vertical lines indicate the 10th and 90th percentiles. Dots indicate outliers outside the 10th and 90th percentiles. n.s., not significant; ***$p < 0.001$.

different requirements for lesson plans within our sample (Table 3), so these criteria will not be considered. Among the remaining criteria, the three criteria with the largest effect sizes were *Choice of topic* (C2; $d = 0.62$), *Progression of competence development* (S2; $d = 0.66$), and *Adaptive teaching* (I7; $d = 1.08$). To provide qualitative insights into PSTs' and TSTs' lesson planning competence and highlight significant differences in their lesson planning, we purposefully selected three PSTs who achieved level 1 and three TSTs who achieved level 2 for the abovementioned criteria. These cases can be regarded as representative examples, not holistically for all criteria on levels 1 or 2, respectively, but at least for the three selected criteria.

6.7.1 | Choice of topic

PST 22 planned a lesson about genetic fingerprints. Students are intended to describe how criminologists use DNA to determine a connection between biological evidence (e.g., hair, skin cells) and a suspect. To justify the relevance of this topic, PST 22 refers to its relevance in science, in society, and for students:

> Over the last decades, molecular genetics has become an important scientific discipline. Researchers have developed methods in many fields, such as pharmacy, criminology, and genealogy. […] Particularly, genetic fingerprints have raised the interest of many people beyond scientists, for example, regarding paternity tests and the search for missing relatives. […] Such applications affect students' lives and the media they consume. Therefore, this topic is relevant for students. (pp. 9–11)

Even though PST 22 attempts to use a topic relevant to the real world and connect to students' interests, they seem to use these elaborations primarily to justify the topic rather than as a starting point for the instruction. In the planned lesson, PST 22 solely focuses on the relevance of genetic fingerprints in science without explicitly adapting to his/her students' funds of knowledge and the topic's relevance for his/her particular students.

In contrast, TST 43 planned a lesson about the neuronal and hormonal regulation of stress. The students are intended to create a diagram illustrating how the human body responds to stressful situations. To justify the relevance of this topic, TST 43 elaborates on their students' current situation at school in general and during the COVID-19 pandemic in particular:

> School is not always fun for students. Particularly during the ongoing pandemic, students face extraordinary challenges. The school was a place to learn *and* interact socially with peers, make friends, and so on. […] Many leisure activities (meeting in public spaces, sports, etc.) are still impossible apart from school. However, students still have to take exams and are still required to perform well, which I assume to be even more stressful than in regular times. Due to that, it is essential for me to deal with stress from a biological perspective and have my students reflect and discuss how they could cope with stress in their individual lives. (p. 6)

TABLE 7 Descriptive statistics for both the RALP scoring and the holistic quality assessment.

		RALP scoring				Holistic quality assessment				
	n	Mean	SD	Min	Max	Very good	Good	Satisfactory	Sufficient	Poor
PSTs	36	20.5	5.0	11	30	10	15	7	3	1
TSTs	64	27.6	5.5	15	39	13	14	17	15	5

Note: The range of possible RALP scores is 0–48.

Abbreviations: Max, maximum value; Min, minimum value; PSTs, preservice science teachers; RALP, rubric to assess science lesson plans; SD, standard deviation; TSTs, trainee science teachers.

Based on this elaboration, TST 43 plans to start the lesson by pretending to take a test on the contents of the previous lesson that took place virtually. TST 43 plans to stop this simulated test situation after 3 min and ask students how they felt during this unannounced test. Based on their initial thoughts and after reading info texts on the physiological regulation of stress, students are intended to collect stressful situations in their personal lives, classify them, and develop ideas for reducing stress.

In summary, these two examples illustrate the typical differences in how prospective science teachers try to bridge the gap between the subject matter that needs to be taught on the one hand and their students' experiences on the other hand: While PST 22 justifies the choice of the topic with general arguments (i.e., the curriculum, relevance in science, society), TST 43 explicitly tries to connect to their students and attempts to make biology relevant to them. Most importantly, unlike PST 22, TST 43 explicitly connects to their students' experiences and tailors the lesson to their needs.

6.7.2 | Progression of competence development

PST 34 planned a lesson about the human circulatory system. In their analysis of their students' current level of competence, PST 34 claims:

> All students can describe the structure and function of the heart as well as the difference between veins and arteries. […] They can explain the connection between pulse and respiration. […] Only some students can describe the circulatory system as a whole. […] (p. 3)

Even though PST 34 explicitly states that all students already can describe and explain the structure and function of the circulatory system, the students are intended to read a text and solve tasks concerning the heart, veins, and arteries, which can be regarded as a repetitive exercise rather than a progression. Beyond that, the text contains information about systemic and pulmonary regulation interplay, which is new knowledge for most students. In the final task, students are provided a schematic illustration of the human body showing the heart, brain, arteries, and veins colored red and blue, respectively. Students are intended to assign given technical terms (e.g., aorta, pulmonary artery) to the illustration. It does not become clear how far this exceeds the students' current level of competence. By now, only some students seem able to describe the circulatory system as a whole. Learning that would require a more holistic approach, for example, tracing a blood cell from the heart through the body. However, this is not part of the planned lesson.

In contrast, TST 2 planned a lesson about sustainability by evaluating a set of options for action to reduce one's ecological footprint. The criteria for such an evaluation (e.g., economic costs, health, social justice) were developed in the previous lesson. To justify the need for this lesson, TST 2 analyzes their students' current level of competence:

> The students have already practiced reflecting on and evaluating phenomena in biological/ecological contexts. [They] have applied this procedure to a discussion in society and politics about reducing carbon dioxide emissions. [They] have yet to reflect and analyze their behavior and impact on the environment. (p.3)

Based on this analysis, TST 2 has students calculate their ecological footprint and develop ideas for reducing their direct or indirect carbon dioxide emissions. Finally, students are intended to apply the criteria to discuss and evaluate the suitability of their ideas. Thus, they must reflect on their behavior critically, weigh the pros and cons of their personal lives and discuss concrete options for deciding to reduce their ecological footprint. This is a

progression since TST 2 stated that, so far, their students have reflected on the consequences of political or economic decisions on ecology, which is relatively abstract.

In summary, these two examples illustrate the typical differences in how prospective science teachers adapt to their students' current level of competencies: While PST 14 provides a mixture of repetitive and new elements without explicitly adapting to their students' current needs, TST 2 explicitly builds upon the criteria that were developed in the previous lessons and transferred the procedure of an ethical evaluation of political or economic behavior to each student's behavior. In doing so, TST 2 explicitly adapts to their students' current needs and helps them develop their competencies.

6.7.3 | Adaptive teaching

PST 9 planned a lesson about Mendel's Laws of Heredity by crossing a black and a brown dog, resulting in a uniform generation of dogs with black and brown fur. Most students can "develop crossing schemes for Mendel's Laws […] and describe them with correct technical terms." However, PST 9 knows some students might even struggle with it. To help them out, they prepared a worksheet already containing the correct crossing scheme so that students "have more time to reread the text and deal with the hypotheses [that were collected at the beginning of the lesson]." In doing so, PST 9 provides additional help for low-performing or relatively slowly-working students. However, it might need to be reflected in how far this approach might go too far because providing solutions is more than helping students find their way to a solution.

In contrast, TST 30 planned a lesson on the symbiotic relationship between mycorrhizal fungi and plant roots by creating an infographic using the example of truffles. In their analysis of the students' current needs, TST 30 claims that the students form a highly heterogeneous group concerning their ability to illustrate textual or numerical information: While "some students still struggle to read tables and decide which type of diagram would be most suitable to visualize the data," the top-performing students can read and understand complex texts and discuss the pros and cons of various options for visualization. To adapt to these heterogenous needs in the class, TST 30 uses two instructional approaches: First, low-performing students are provided aid cards to have students think about how to illustrate the information adequately. For instance, one of the pieces of advice is: "Recall our recent discussion on the exam question: Wouldn't it be necessary to use two y-axes in the diagram? Define which data are dependent and which are independent variables and assign them to the x- or y-axis." Second, the high-performing students are provided an additional task because they usually work more quickly than the rest of the class. They are requested to "include and explicitly name biotic (e.g., boars, pathogens) and abiotic factors (e.g., soil, the temperature during winter) that influence the growth of truffles" and thus illustrate the textual and numerical information rather than in the form of a concept map.

In summary, these two examples illustrate typical differences in how prospective science teachers adapt to their students' current level of competencies: While PST 9 at least offers additional help for low-performing students, TST 30 plans to provide tasks and help for both the low- and the high-performing students.

7 | DISCUSSION

There is a lack of high-quality instruments to assess science teachers' lesson planning competence, in general, and their ability to write proper lesson plans, in particular (Krepf & König, 2022). This study aimed to build an argument for the validity of the interpretations of the RALP scores by considering multiple sources of validity evidence (AERA, APA, & NCME, 2014; Brookhart, 2018; Jonsson & Svingby, 2007) and applying this novel instrument to a sample of $N = 100$ prospective science teachers to gain further insights into the development of lesson planning competence. We regard the employed explanatory sequential mixed-methods design (QUAN → qual; Creswell & Plano

Clark, 2018) as one of the strengths of this study: On the one hand, the relatively large sample size allowed for a quantitative analysis that yielded significant findings concerning the validity of the RALP score interpretation. However, the quantitative analysis alone would not provide further insights into prospective science teachers' lesson planning. On the other hand, the qualitative analysis of three pairs of lesson plans showed challenges that PSTs and TSTs meet in the process of lesson planning on different levels of quality. However, the qualitative analysis alone would barely suffice to claim that the findings are generalizable, at least to a certain degree. The combination of the quantitative and the qualitative approach made it possible to apply the RALP validly and to purposefully select the three most fruitful criteria for comparing PSTs' and TSTs' lesson plans.

To the best of our knowledge, the RALP is currently the only available broadly applicable scoring rubric in science education that can be used to assess lesson planning competence. Unlike other existing rubrics in science education (Table 1: Enugu & Hokayem, 2017; Forbes & Davis, 2010; Goldston et al., 2013), the RALP can be applied not only to inquiry lessons but to all kinds of science lessons. Moreover, in contrast to Jacobs et al.'s (2008) rubric, the RALP is theoretically framed and connects to PCK-based research designs. Thus, even though we solely analyzed biology lesson plans, the RALP might also apply to chemistry or physics lesson plans. In the following section, we will first reflect on a couple of relevant limitations in this study and then discuss key characteristics of the RALP with regard to the abovementioned validity argument (Table 2). For the sake of comprehensibility, these characteristics will not be discussed in chronological order but sorted by the three vertices of Pellegrino et al.'s (2016) *Assessment Triangle* (Figure 2).

7.1 | Limitations

Overall, the RALP's application to $N = 100$ lesson plans limited the possibility of statistical analyses. It was not manageable to obtain a more considerable amount of lesson plans written by PSTs from the university as, by now, the criteria are used in the authors' courses so that PSTs are trained to consider them. Consequently, results would be skewed if we had analyzed a sample trained to apply the RALP. Hence we analyzed lesson plans from 2019 to 2021, which were written under identical circumstances by PSTs who did not know the RALP yet. Most importantly, the present study is essentially qualitative and thus seems to have a large sample size compared to other qualitative studies analyzing written lesson plans (e.g., $N = 10$, Chizhik & Chizhik, 2018; $N = 22$, Kademian & Davis, 2018; $N = 10$, Morine-Dershimer, 1979; $N = 18$, Zaragoza et al., 2021).

Moreover, the analysis of these lesson plans also implicates several methodological challenges. For instance, we did not influence the criteria and formal regulations the TSTs considered when writing the lesson plans. Consequently, the word counts of TSTs' lesson plans ranged between approximately 1800 and 7800 (Großmann & Krüger, 2022c), affecting the probability that all 24 RALP criteria were equally considered. A more controlled setting where PSTs and TSTs write lesson plans might seem more appropriate. Thus, we would ensure we capture the prospective teachers' $ePCK_p$ or pPCK, respectively. Since the lesson plans analyzed in this study were part of formal examinations, PSTs and TSTs might have received feedback from colleagues or friends, leading us to capture collective PCK instead (Alonzo et al., 2019) rather than exclusively pPCK. However, writing a lesson plan is time-consuming and challenging to integrate into a typical university course. As described above, Kang (2017) suggested that researchers in the field of lesson planning should not rely on controlled research designs but consider teachers planning processes or products captured in natural environments. To fill this gap, we decided to focus on original lesson plans. We regard the ecological validity (Wegener & Blankenship, 2007) of our analysis as a crucial feature in the development and evaluation process of the RALP. Nevertheless, it might have been helpful to collect demographic information about the participants or information about the schools they were teaching at to get a complete impression of the circumstances under which the participants were planning and teaching.

Recently, there has been a shift from developing scoring rubrics for grading purposes (summative assessment) to supporting the target groups in their learning processes (formative assessment; Panadero & Jonsson, 2013;

Panadero et al., 2018). Thus, this study might be regarded as a shift back to a past time since we focused on the summative perspective. However, this study can also be regarded as the foundation for the further development of the RALP: As a first step, it needed to be investigated whether the RALP covers relevant aspects of lesson planning (RQ 1.1), whether it can distinguish quality levels (RQ 1.5), and whether it is instructional sensitive (RQ 1.6). Based upon that, further evidence should be collected, particularly with regard to formative and self-assessment (e.g., *validity evidence based on response processes*; AERA, APA, & NCME, 2014). As Brookhart (2018) suggests, future studies should consider how the target group uses a scoring rubric as a learning tool.

7.2 | Evaluation of the plausibility of the validity argument

To ensure that the scores of the RALP can be interpreted validly, empirical evidence is needed (AERA, APA, & NCME, 2014; Kane, 2013; Pellegrino et al., 2016), which is often not considered in research on scoring rubrics (Reddy & Andrade, 2010) and with regard to scoring rubrics on lesson planning (Table 1). As the underlying validity argument, we defined that the RALP scores are interpreted as indicators of lesson plan quality. Overall, the RALP covers essential aspects of lesson planning and is, first and foremost, able to discriminate between high- and low-quality lesson plans. Therefore, this section will take step III of building a validity argument (Figure 2) and discuss the empirical findings for all three vertices of the *Assessment Triangle* (Table 2; Pellegrino et al., 2016).

7.2.1 | Evidence for *cognitive validity*

Proceeding from the assumption that lesson planning requires professional knowledge in general (Zaragoza et al., 2021) and PCK in particular (Alonzo et al., 2019; Carlson et al., 2019; Großmann & Krüger, 2022c; Vogelsang et al., 2022), it appears reasonable to operationalize lesson planning competence in terms of PCK. Figure 1 shows that the 24 cognitive demands during lesson planning built upon König et al.'s (2021) CODE-PLAN model can be assigned to four of the five PCK components and their interconnections (Park & Oliver, 2008). In addition, as described by Großmann and Krüger (2022a), all criteria and performance levels build upon theoretical and empirical research. Thus, it can be assumed that the RALP sufficiently relates to the cognitive construct (*Validity evidence based on test content*; AERA, APA, & NCME, 2014). This was also confirmed by the expert discussions (RQ 1.1). Based on their feedback, we rechecked and improved the assignment of criteria to the pentagon model, the order of the criteria within the RALP, and the specificity and appropriateness of the performance level descriptions.

7.2.2 | Evidence for instructional validity

Among the 29 initial criteria rated by the teacher educators (RQ 1.1), only two were considered somewhat irrelevant: the foundation of justifications with theoretical and empirical literature and considering students' conceptions (Table 4). We accepted the suggestion to include justifications in the performance level descriptions of selected RALP criteria and not provide an individual criterion. Regarding students' conceptions, some teacher educators argued that it is unnecessary to consider them in each biology lesson, so they rated their relevance rather poorly. The neglect of students' conceptions in trainee biology teachers' lesson plans was recently described by Großmann and Krüger (2022b). However, students' conceptions play a significant role in science education (e.g., Lucero et al., 2017; Otero & Nathan, 2008) since they would allow educators to adapt their instruction to students' thinking in terms of the constructivist learning theory (Duit & Treagust, 2012). Due to that, and since the criteria are also meant to trigger prospective teachers to consider aspects they otherwise might miss during lesson planning, we decided to keep students' conceptions as a criterion in the RALP. Two teacher educators suggested selecting relevant technical terms as

another criterion. This is also an issue in science education research, showing that in-service biology teachers' extensive use of technical terms affects students' achievements and situational interests (Dorfner et al., 2020). We concluded that this aspect is relevant for prospective teachers and thus added it to the RALP.

Regarding the analysis of validity evidence based on the internal structure (RQ 1.3, RQ 1.4), the findings align with our assumptions. As expected, due to different requirements (Table 3), PSTs barely contextualized their lesson into a unit (O1, O2), whereas TSTs barely considered students' conceptions (S3) in their lesson plans (Table 6). More specifically, the intended learning outcomes often do not mediate prospective teachers' thinking about their students' needs (S1). Particularly, assessment at the end of a lesson is often not fully aligned with the intended learning outcome (A2), a frequently reported issue in lesson planning (e.g., Chizhik & Chizhik, 2018; Weitzel & Blank, 2020). Hence, the findings are also in line with other studies. Moreover, prospective teachers struggled to describe and justify the tasks they planned to give (I4). Among the many aspects to consider, planning tasks represent one of the essential cognitive demands (König et al., 2021). Tasks should be aligned with the intended learning outcome and match the students' current abilities to facilitate a meaningful learning process. Developing high-quality tasks is challenging for prospective teachers, mainly because it is also related to the phasing of a lesson (Kang et al., 2016). Thus, it is one of the critical decisions in the lesson planning process.

The psychometric analysis provides evidence for a one-dimensional model (Table 5). Thus the competence to write good lesson plans should not be separated into two or more latent variables. We interpret this as an indicator that the RALP criteria reflect the interconnectedness of PCK components (Figure 1) rather than distinct aspects. In contrast, König et al. (2021) suggested that lesson planning competence consists of six skills (content transformation, task creation, adaptation to student learning dispositions, clarity of learning objectives, unit contextualization, and phasing) that are needed to master six significant cognitive demands in lesson planning. Their confirmatory factor analysis showed a good model fit for a six-dimensional model and a poor model fit for a one-dimensional model. Thus, we would not argue that lesson planning competence generally cannot be specified in terms of latent variables. We only conclude that for our sample of $N = 100$ prospective science teachers, the RALP reflects the intended unidimensionality instead of a reasonably assumed multidimensionality. Possibly, the ability to meet König et al.'s (2021) six cognitive demands might indeed form separable components of lesson planning competence. In contrast, the related PCK components investigated in this study develop their potential if they are interconnected, thus leading to a one-dimensional model. Based on the assumption that the underlying construct is unidimensional, Cronbach's α (0.78) was calculated as a measure of internal consistency and is sufficiently high in this sample. Still, it should be noted that Cronbach's α is affected by the number of criteria (Taber, 2018), so this value might result from many criteria ($n = 24$) sharing some variance with others. In addition, the Wright map (Figure 4) suggests that prospective science teachers' ability to address the RALP criteria is normally distributed; most criteria match their ability. Hence the performance expectations were neither too demanding nor under-demanding for our sample. While the TSTs barely considered students' conceptions, and thus this criterion formed an outlier in the Wright map, the alignment of teachers' horizon of expectations to their students' current levels of competence (I6) might be considered a superfluous criterion.

However, since this analytic rubric is also intended to make expectations of good lesson plans transparent and to be used for formative feedback, we would keep these criteria. Most criteria exceed the threshold of 0.30 (Vaus, 2014) and can thus be regarded as sufficiently discriminatory. The remaining four criteria (O4, S3, S5, I4; Supporting Information: Appendix A) barely help to discriminate PSTs' and TSTs' performances (Table 6). Suppose the RALP is intended for summative assessment purposes. In that case, some highly discriminatory criteria might receive a higher weight in the scoring process, indicating that these criteria contribute more to a good lesson plan than others.

7.2.3 | Evidence for *inferential validity*

The RALP allows a stable (RQ 1.2a) and intersubjectively comprehensible (RQ 1.2b) scoring procedure. As stated above, we intend the RALP to be used by teacher educators for grading purposes and by TSTs as a learning and

self-assessment tool (Brookhart, 2018; Krebs et al., 2022; Panadero et al., 2018). For that purpose, we suggest providing at least one coded lesson plan in addition to the scoring rubric so that the target group gets an idea of how the rubric can be applied and which parts of a typical lesson they mainly need to focus on. Additional explanations and definitions might also be necessary to clarify what a good intended learning outcome (O3) is or when an activity can be regarded as appropriate (I3), which were two criteria with problematic κ values. Moreover, even though it is relatively uncommon to investigate intrarater agreement in research on scoring rubrics (Brookhart, 2018; Jonsson & Svingby, 2007), we argue that the high κ values are an essential prerequisite for the use of the RALP and should therefore be calculated when developing a scoring rubric.

Most importantly, we found evidence for the "predictive validity" of the RALP: The excellent correlations between the RALP scores and the teacher educator's holistic quality assessment and the large effect sizes between quality levels (RQ 1.5; *Validity evidence based on relations to other variables*; AERA, APA, & NCME, 2014) indicate that the RALP scores do indeed reflect lesson plan quality. The RALP can discriminate quality levels for the whole subsamples and between the subgroups of holistic quality assessment (Figure 5). In line with Tomas et al. (2019), we argue that analytic and holistic approaches should not be regarded as mutually exclusive but as complementary possibilities for substantiating interpretations made from analytic or holistic rubrics. However, if the RALP is intended for grading, it should be considered to add weights to the criteria, as executed by Jacobs et al. (2008). All 24 criteria contribute equally to the final score, which might skew the results. Probably, criteria referring to the intended learning outcome as the most crucial reference point for making planning decisions during lesson planning (Drost & Levine, 2015) might be more important than a justification of the choice of topic, so this might be taken into account in the scoring process by adding weights. Another option would be to use the relevance assessment (Table 4) to weigh the most relevant criteria (e.g., >2.70) three times, the criteria of medium relevance (e.g., 2.30–2.69) two times, and the minor relevant criteria (e.g., <2.30) one time. Hence, the scores 0, 1, and 2 for each criterion must be multiplied by 1, 2, or 3, respectively. Thus, one could ensure that the criteria found to be most relevant (e.g., formulating a precise intended learning outcome) contribute more to the overall score than criteria of minor relevance (e.g., considering students' conceptions). Another option would be that teacher educators who wish to use the RALP in their courses determine weights based on their individual preferences.

Moreover, the RALP consists of only three performance levels. Most published analytic scoring rubrics have four or five levels (Brookhart, 2018). In developing the RALP, we had to weigh the pros and cons of increasing or decreasing the number of performance levels. On the one hand, more performance levels would increase the instrument's sensitivity as it would detect quality differences on a more fine-grained level. While the holistic quality assessment is relatively fine-grained (1.0, 1.3, …, 4.7, 5.0), the three performance levels only allow a coarse-grained analysis. On the other hand, the performance levels should be selective to ensure interrater reliability. Developing more than three distinct performance level descriptions was difficult for many criteria. Consequently, we accepted a lower sensitivity in favor of clarity and unambiguity and decided to provide three performance levels for each criterion.

Apart from the development process, even the top-rated TSTs who achieved the highest RALP scores could not even come close to the maximum score of 48, meaning that lesson plans can be of high quality even without considering every criterion. This might be another reasonable argument to add weight to the criteria and thus indicate that specific criteria contribute more to the overall lesson plan quality than others. However, we would not delete any criteria from the RALP since they were considered relevant by teacher educators (RQ 1.1). Moreover, neither the PSTs nor the TSTs in this sample explicitly practiced applying the 24 criteria because they had not existed yet when they wrote their lesson plans. Teacher educators that use the RALP in their courses should introduce the 24 criteria to ensure transparency. We expect that under such circumstances, the maximum score of 48 will be achievable by high-performing students.

Moreover, by our expectations, the more experienced TSTs scored significantly higher than the PSTs (RQ 1.6; Table 7; Figure 5). Even though we did not conduct a longitudinal but a cross-sectional study, our findings connect to previous research on the development of lesson planning competence during teacher training (Backfisch

et al., 2020; König et al., 2021, 2022; Mutton et al., 2011; Vogelsang et al., 2022; Westerman, 1991). TSTs gained much more field experience and professional knowledge during their induction phase than PSTs during their studies (Figure 3). Possibly, this facilitates the integration of different PK, CK, and PCK aspects during lesson planning (Stender et al., 2017; Zaragoza et al., 2021) and results in higher scores on the RALP. Another fruitful idea for teacher educators might be to distinguish *lesson plan analysis* and *lesson planning* and scaffold the application of professional knowledge to both practices in a long-term learning process (Zaragoza et al., 2023): Before PSTs engage in planning their lessons, teacher educators might provide a scaffold and have PSTs analyze simulated or authentic lesson plans. Thus, they would gradually be introduced to the complex nature of lesson planning. For the first step in Zaragoza et al.'s (2023) suggestion—the lesson plan analysis—the 24 RALP criteria might be used. After PSTs have learned to evaluate lesson plans, they start planning their own lessons and might use the RALP criteria as guidelines.

Finally, it should be noted that PSTs' and TSTs' lesson plans were written under different conditions: While it is only the third lesson plan that PSTs usually write in their teacher training, which is only one of many examinations that PSTs have to complete during Master's studies, TSTs have gained far more practice during their induction phase and can focus on the final examination which is of unique importance (Figure 3). Hence, as König et al. (2021) argue, expectations regarding lesson plan quality are much higher for TSTs' final examination. Therefore they might put more effort into it than PSTs, which might explain the large effect.

7.3 | Differences between PSTs' and TSTs' lesson plans

Our study adds to the research attempts in the field of lesson planning in science education mentioned above (Table 1) by applying the RALP to original lesson plans and providing qualitative insights into how prospective science teachers plan lessons and justify their decisions. The sample was large enough to claim that the following observations can be generalizable, at least for prospective biology teachers in Berlin. While novices need help to interconnect their professional knowledge in the process of lesson planning, expert teachers use their knowledge to explicitly adapt to their student's needs (Westerman, 1991). Our analysis sheds light on the differences between PSTs and TSTs.

In general, the more experienced TSTs focus clearly on their particular students and tailor the planned instruction and their justifications to their student's needs. The comparison between PST 22 and TST 43 illustrates a typical pattern found in many lesson plans: While TSTs use the analysis of the biological subject matter and their students' needs to plan a lesson, PSTs often elaborate in detail on rather irrelevant information. Since interconnecting these different knowledge bases with each other is challenging, particularly for novices (Westerman, 1991; Zaragoza et al., 2021), it seems to be necessary for teacher educators to provide support for lesson planning (Karlström & Hamza, 2021) and highlight for what purpose a lesson plan needs to be written. This might help PSTs to focus on essential aspects of teaching. Regarding the choice of topic, PST 22 justified the relevance with rather broad arguments, referring to society, for example. As König et al. (2020) point out, preservice teachers frequently use generic statements in lesson planning. Instead, it might be necessary to focus on subject-specific arguments to highlight how particular planning decisions help students learn science.

Regarding instructional planning, the significant difference between PSTs and TSTs in this sample is the degree to which the prospective science teachers adapt to their students' needs and plan lessons explicitly tailored to them (Westerman, 1991). Unlike experienced teachers, PSTs still lack mental plans (Borko & Livingston, 1989), which might explain their difficulties in building a lesson that connects to their students' current level of competencies. In their lesson plan, TST 2 connected to the previous lesson and made the transitions between the phases within the lesson transparent. The lesson is contextualized in a unit, resulting in a long-term learning process that entails a stepwise and progressive sequence of lessons to help students develop their competencies. The PSTs in this sample were not required to plan a whole unit (Table 3). Since unit contextualization is one of the significant cognitive

demands in lesson planning (König et al., 2021), teacher educators might reflect on whether teaching how to plan long-term learning processes needs to be considered early on in teacher education.

Regarding the adaptation to the students' needs, the RALP might help PSTs to broaden their perspective on their students: It is worth emphasizing that even the PSTs often consider how to help the low-performing students achieve the intended learning outcome, which can be regarded as an encouraging result. TST 30 adapts to the high-performing students' abilities, indicating a differentiated view of the class and the capacity to use instructional strategies to consider individual differences. There is no convincing reason why PSTs could not use such strategies for high-performing students. In that sense, the RALP might encourage PSTs to take low- and high-performing students into account, and thus, prospective science teachers might use it as a learning tool.

8 | CONCLUSION AND OUTLOOK

This study pursued two objectives: First, we aimed to evaluate the quality of the RALP considering multiple sources of validity evidence (AERA, APA, & NCME, 2014). Second, we aimed to apply the RALP to an ecologically valid sample of prospective science teachers to gain qualitative insights into their strengths and weaknesses. As suggested by Scherer (2017), referring to Pellegrino et al.'s (2016) *Assessment Triangle* (Figure 2) facilitates building a sound and persuasive validity argument since it automatically requires considering different participants (e.g., science teacher educators, PSTs, TSTs). Overall, various pieces of empirical evidence for validity were collected, and thus, a persuasive validity argument was built. In particular, the RALP covers relevant aspects of lesson planning and can discriminate lesson plan quality and can thus be regarded to be a valuable instrument for (1) teacher educators and (2) researchers: (1) Teacher educators can use the RALP for both grading and feedback purposes in teacher training. In doing so, teacher educators could make their expectations transparent and refer to an objective set of criteria instead of their own, maybe partially subjective criteria. Since the RALP is an analytic scoring rubric, feedback would be highly differentiated and help prospective science teachers to grasp the next step in their learning process. Therefore, we recommend not displaying the overall score but only focusing on single criteria (Panadero & Jonsson, 2020). (2) Researchers in science education might use the RALP in study designs that include the analysis of lesson plans, for example, in lesson studies (e.g., Juhler, 2018). Since the RALP can be used for both qualitative analyses based on the performance level descriptions and for quantitative analyses based on the overall score, we argue that this instrument might contribute to research in science education in various ways since it allows valid interpretations. The qualitative analysis of selected typical cases has shed light on significant differences between PSTs and TSTs, particularly regarding their ability to adapt to their particular students' needs. Presenting some of the RALP criteria to PSTs might help them consider aspects they might have missed otherwise. Moreover, the criteria might facilitate focusing on relevant aspects for instruction instead of detailed elaborations on aspects of minor relevance for the class. However, it should be noted that the RALP was developed in the context of German teacher education. As mentioned above, lesson plans consist of approximately 10 pages containing analyses, descriptions, and justifications. This concept of lesson plans differs from those common in other countries, where lesson plans consist of three pages at most. Consequently, if teacher educators or researchers intend to apply the RALP to their contexts, we recommend ensuring that the RALP criteria meet the local requirements. If PSTs are not required to justify their planning decisions, they cannot reach a high score. However, this would result from an inappropriate assessment procedure rather than the preservice teachers' failure. We suggest adjusting the lesson plan requirements to apply the RALP appropriately in these cases. The other option would be to modify the RALP according to the locally different lesson plan requirements. However, this might have undesired effects on certain validity aspects. For instance, deleting items might affect *validity evidence based on test content*, as critical, theoretically reasoned, and empirically tested criteria would be abandoned.

Since "the validation process never ends, as there is always additional information that can be gathered to more fully understand a test and the inferences that can be drawn from it" (AERA, APA, & NCME, 2014, pp. 21–22), we suggest investigating further aspects of validity that are of significant concern for the quality of the RALP. First, the assumption that the RALP applies to all sciences must be investigated. Moreover, bearing in mind that the RALP is not only intended to be used as a grading tool for summative assessment purposes but also for formative assessment (Panadero & Jonsson, 2020) and self-assessment purposes (Krebs et al., 2022), the following research desiderata (i.e., open questions for further research) particularly relate to the interpretation vertex of the *Assessment Triangle* (Figure 2; Pellegrino et al., 2016): If measuring prospective teachers' PCK (e.g., using the PCK-in biology inventory; Großschedl et al., 2019), is there evidence of convergent validity, indicating that the RALP does indeed measure PCK (*Validity evidence based on relations to other variables*)? How do prospective teachers engage with the RALP when applying it to their lesson plans? Can they improve a lesson plan after self-assessment (Krebs et al., 2022; Scherer, 2017; *Validity evidence based on response processes*)? If the RALP is applied by a teacher educator, by a peer student (peer assessment), or by students themselves (self-assessment) and the authors of the lesson plans are requested to improve their lesson plan afterward (Ozogul et al., 2008): Which of these three types of application helps prospective teachers to improve their lesson plans most? In an experimental-control-group-design, do prospective teachers who were introduced to the RALP write better lesson plans in terms of holistic quality assessment than those who do not know the RALP criteria beforehand (*Validity evidence based on the consequences of testing*)? Finally, we recommend shedding light on the underexplored relationship between lesson planning and instruction. In what way do the RALP scores predict the quality of classroom teaching?

ACKNOWLEDGMENTS

The project K2Teach (Know how to teach) is part of the "Qualitätsoffensive Lehrerbildung," a joint initiative of the German Federal Government and the Länder, which aims to improve the quality of teacher training. The programme is funded by the Federal Ministry of Education and Research (Grant Number 01JA1802). The authors are responsible for the content of this publication. Furthermore, the authors would like to thank the Berlin Senate Department for Education, Youth, and Family for providing lesson plans in an anonymized form. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The lesson plans were anonymized, digitized, and analyzed exclusively to answer the scientific research questions. The trainee science teachers' lesson plans were provided, thankfully, by the Berlin Senate Department for Education, Youth, and Family under the condition that they are not published or disseminated. The copyright remains with the biology teachers who have written the lesson plans.

ORCID

Leroy Großmann http://orcid.org/0000-0001-7635-1737
Dirk Krüger http://orcid.org/0000-0003-0999-4382

ENDNOTES

[1] It has not escaped our notice that there are persistent concerns in PCK research on the varying conceptualization of PCK, on measuring PCK, and on the comparability of research findings that might be affected by those different understandings and measures (e.g., Chan & Hume, 2019; Settlage, 2013). Consequently, caution is advised when correlating PCK to other variables, such as cognitive activation. Nevertheless, the relevance of PCK for the quality of teachers' practice is beyond doubt.

[2] Note that Kane (2013, p. 10) differentiated between the *interpretation/use argument*, "that specifies what is being claimed in the interpretation and use," and the *validity argument*, "[…] which provides an evaluation of the proposed [interpretation/use argument]." Agreeing with Newton's (2013) concluding remarks at the end of his critical analysis of the expediency of Kane's (2013) terminological differentiation, we assume there is only one argument, the *validity argument*, which is the term we will adhere to in this paper.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Psychological Association.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*(1), 3–14. https://doi.org/10.1016/0304-4076(81)90071-3

Alonzo, A. C., Berry, A., & Nilsson, P. (2019). Unpacking the complexity of science teachers' PCK in action: Enacted and personal PCK. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 271–286). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_12

Andrade, H. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*, *53*(1), 27–31.

Australian Institute for Teaching and School Leadership. (2018). *Australian professional standards for teachers*. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/australian-professional-standards-for-teachers.pdf

Backfisch, I., Lachner, A., Hische, C., Loose, F., & Scheiter, K. (2020). Professional knowledge or motivation? Investigating the role of teachers' expertise on the quality of technology-enhanced lesson plans. *Learning and Instruction*, *66*, 101300. https://doi.org/10.1016/j.learninstruc.2019.101300

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Wright maps: First steps. In W. J. Boone, J. R. Staver, & M. S. Yale (Eds.), *Rasch analysis in the human sciences* (pp. 111–136). Springer. https://doi.org/10.1007/978-94-007-6857-4_6

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal*, *26*(4), 473–498.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*(22), 1–12. https://doi.org/10.3389/feduc.2018.00022

Campbell, T., Gray, R., Fazio, X., & van Driel, J. (2022). Research on secondary science teacher preparation. In J. A. Luft & M. G. Jones (Eds.), *Handbook of research on science teacher education* (pp. 97–118). Routledge. https://doi.org/10.4324/9781003098478-10

Carlson, J., Daehler, K. R., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K. K. H., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze_Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., … Wilson, C. D. (2019). The refined consensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper & A. Borowski, (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77–92). Springer. https://doi.org/10.1007/978-981-13-5898-2_2

Chan, K. K. H., & Hume, A. (2019). Towards a consensus model: Literature review of how science teachers' pedagogical content knowledge is investigated in empirical studies. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 3–76). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_1

Chan, K. K. H., Rollnick, M., & Gess-Newsome, J. (2019). A grand rubric for measuring science teachers' pedagogical content knowledge. In A. Hume, R. Cooper & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 253–271). Springer. https://doi.org/10.1007/978-981-13-5898-2_11

Chizhik, E. W., & Chizhik, A. W. (2018). Using activity theory to examine how teachers' lesson plans meet students' learning needs. *The Teacher Educator*, *53*(1), 67–85. https://doi.org/10.1080/08878730.2017.1296913

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research (international student edition)* (3rd ed.). Sage.

Darling-Hammond, L., Newton, S. P., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability*, *25*(3), 179–204. https://doi.org/10.1007/s11092-013-9163-0

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

Dorfner, T., Förtsch, C., & Neuhaus, B. J. (2020). Use of technical terms in German biology lessons and its effects on students' conceptual learning. *Research in Science & Technological Education*, *38*(2), 227–251. https://doi.org/10.1080/02635143.2019.1609436

Drost, B. R., & Levine, A. C. (2015). An analysis of strategies for teaching standards-based lesson plan alignment to preservice teachers. *Journal of Education*, *195*(2), 37–47.

Duit, R., & Treagust, D. F. (2012). How can conceptual change contribute to theory and practice in science. In B. J. Fraser, K. Tobin, & C. McRobbie (Eds.), *Second international handbook of science education* (pp. 107–118). Springer.

Enugu, R., & Hokayem, H. (2017). Challenges pre-service teachers face when implementing a 5E inquiry model of instruction. *European Journal of Science and Mathematics Education*, *5*(2), 178–209.

Forbes, C. T., & Davis, E. A. (2010). Curriculum design for inquiry: Preservice elementary teachers' mobilization and adaptation of science curriculum materials. *Journal of Research in Science Teaching*, *47*(7), 820–839. https://doi.org/10.1002/tea.20379

Förtsch, C., Werner, S., von Kotzebue, L., & Neuhaus, B. J. (2016). Effects of biology teachers' professional knowledge and cognitive activation on students' achievement. *International Journal of Science Education*,*38*(17), 2642–2666. https://doi.org/10.1080/09500693.2016.1257170

Gess-Newsome, J. (2015). A model of teacher professional knowledge and skill including PCK: Results of the thinking from the PCK Summit. In A. Berry, P. M. Friedrichsen & J. Loughran (Eds.), *Teaching and learning in science series. Re-examining pedagogical content knowledge in science education* (pp. 28–42). Routledge.

Goldston, M. J., Dantzler, J., Day, J., & Webb, B. (2013). A psychometric approach to the development of a 5E lesson plan scoring instrument for inquiry-based teaching. *Journal of Science Teacher Education*, *24*(3), 527–551. https://doi.org/10.1007/s10972-012-9327-7

Großmann, L., & Krüger, D. (2020). Entwicklung und Anwendung eines Kategoriensystems zur Analyse des fachdidaktischen Wissens angehender Biologie-Lehrkräfte in schriftlichen Unterrichtsplanungen [Development and Application of a category system for the analysis of pre-service biology teachers' pedagogical content knowledge in written lesson plans]. *Erkenntnisweg Biologiedidaktik*, *19*, 21–39.

Großmann, L., & Krüger, D. (2022a). Biologieunterricht erfolgreich planen—ein Kriterienraster zum Schreiben von Unterrichtsentwürfen [Planning biology lessons successfully—a scoring rubric for writing lesson plans]. *SEMINAR—Lehrerbildung und Schule*, *1*, 91–110.

Großmann, L., & Krüger, D. (2022b). Students' conceptions as a neglected perspective in trainee teachers' biology lesson plans. In K. Korfiatis & M. Grace (Eds.), *Contributions from biology education research. Current research in biology education* (pp. 181–193). Springer International Publishing. https://doi.org/10.1007/978-3-030-89480-1_14

Großmann, L., & Krüger, D. (2022c). Welche Rolle spielt das fachdidaktische Wissen von Biologie-Referendar*innen für die Qualität ihrer Unterrichtsentwürfe? [What's the role of trainee biology teachers' pedagogical content knowledge for the quality of their written lesson plans] *Zeitschrift für Didaktik der Naturwissenschaften*, *28*(1), 1–20. https://doi.org/10.1007/s40573-022-00141-w

Großmann, L., & Krüger, D. (2023). Identifying performance levels of enacted pedagogical content knowledge in trainee biology teachers' lesson plans. In G. S. Carvalho, A. S. Afonso, & Z. Anastácio (Eds.), *Contributions from science education research. fostering scientific citizenship in an uncertain world: Selected papers from the ESERA 2021 conference* (pp. 95–116). Springer. https://doi.org/10.1007/978-3-031-32225-9_7

Großschedl, J., Mahler, D., Kleickmann, T., & Harms, U. (2014). Content-Related knowledge of biology teachers from secondary schools: Structure and learning opportunities. *International Journal of Science Education*, *36*(14), 2335–2366. https://doi.org/10.1080/09500693.2014.923949

Großschedl, J., Welter, V., & Harms, U. (2019). A new instrument for measuring pre-service biology teachers' pedagogical content knowledge: The PCK-IBI. *Journal of Research in Science Teaching*, *56*(4), 402–439. https://doi.org/10.1002/tea.21482

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Advances Analytics LLC.

Jacobs, C. L., Martin, S. N., & Otieno, T. C. (2008). A science lesson plan analysis instrument for formative and summative program evaluation of a teacher education program. *Science Education*, *92*(6), 1096–1126. https://doi.org/10.1002/sce.20277

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Juhler, M. V. (2018). Assessment of understanding: Student teachers' preparation, implementation and reflection of a lesson plan for science. *Research in Science Education*, *48*(3), 515–532. https://doi.org/10.1007/s11165-016-9574-2

Kademian, S. M., & Davis, E. A. (2018). Supporting beginning teacher planning of investigation-based science discussions. *Journal of Science Teacher Education*, *29*(8), 712–740. https://doi.org/10.1080/1046560X.2018.1504266

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kang, H. (2017). Preservice teachers' learning to plan intellectually challenging tasks. *Journal of Teacher Education*, *68*(1), 55–68. https://doi.org/10.1177/0022487116676313

Kang, H., Windschitl, M., Stroupe, D., & Thompson, J. (2016). Designing, launching, and implementing high quality learning opportunities for students that advance scientific thinking. *Journal of Research in Science Teaching*, *53*, 1316–1340.

Karlström, M., & Hamza, K. (2021). How do we teach planning to pre-service teachers—A tentative model. *Journal of Science Teacher Education*, *32*, 1–22. https://doi.org/10.1080/1046560X.2021.1875163

Keller, M. M., Neumann, K., & Fischer, H. E. (2017). The impact of physics teachers' pedagogical content knowledge and motivation on students' achievement and interest. *Journal of Research in Science Teaching*, *5*, 586–614.

Koberstein-Schwarz, M., & Meisert, A. (2022). Pedagogical content knowledge in material-based lesson planning of preservice biology teachers. *Teaching and Teacher Education*, *116*, 103745. https://doi.org/10.1016/j.tate.2022.103745

König, J., Bremerich-Vos, A., Buchholtz, C., Fladung, I., & Glutsch, N. (2020). 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., Cammann, F., Bremerich-Vos, A., & Buchholtz, C. (2022). Unterrichtsplanungskompetenz von (angehenden) Deutschlehrkräften der Sekundarstufe: Testkonstruktion und Validierung. *Zeitschrift Für Erziehungswissenschaft*, *25*, 869–894. https://doi.org/10.1007/s11618-022-01113-z

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

Kotthoff, H. G., & Terhart, E. (2013). Teacher education in Germany: Traditional structure, strengths and weaknesses, current reforms. *Scuola Democratica*, *4*(3), 1–9.

von Kotzebue, L. (2022). Beliefs, self-reported or performance-assessed TPACK: What can predict the quality of technology-enhanced biology lesson plans? *Journal of Science Education and Technology*, *31*(5), 570–582. https://doi.org/10.1007/s10956-022-09974-z

Krebs, R., Rothstein, B., & Roelle, J. (2022). Rubrics enhance accuracy and reduce cognitive load in self-assessment. *Metacognition and Learning*, *17*, 627–650. https://doi.org/10.1007/s11409-022-09302-1

Krepf, M., & König, J. (2022). Structuring the lesson: An empirical investigation of pre-service teacher decision-making during the planning of a demonstration lesson. *Journal of Education for Teaching*, 1–16. https://doi.org/10.1080/02607476.2022.2151877

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lucero, M. M., Petrosino, A. J., & Delgado, C. (2017). Exploring the relationship between secondary science teachers' subject matter knowledge and knowledge of student conceptions while teaching evolution by natural selection. *Journal of Research in Science Teaching*, *54*(2), 219–246. https://doi.org/10.1002/tea.21344

Mahler, D., Großschedl, J., & Harms, U. (2017). Using doubly latent multilevel analysis to elucidate relationships between science teachers' professional knowledge and students' performance. *International Journal of Science Education*, *39*(2), 213–237. https://doi.org/10.1080/09500693.2016.1276641

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Ministère de l'Éducation Nationale. (2013). Arrêté du 1er juillet 2013 relatif au référentiel des compétences professionnelles des métiers du professorat et de l'éducation. https://www.legifrance.gouv.fr/loda/id/JORFTEXT000027721614/

Morine-Dershimer, G. (1979). *Teacher plan and classroom reality: The South Bay Study, part IV* (Research Series No. 60). Institute for Research on Teaching College of Education Michigan State University.

Morine-Dershimer, G. (2011). Instructional planning. In J. M. Cooper (Ed.), *Classroom teaching skills* (9th ed., pp. 45–81). Wadsworth CENGAGE Learning.

Moskal, B. M. (2000). Scoring rubrics: What, when and how?. *Practical Assessment, Research, and Evaluation*, *7*, 3. https://doi.org/10.7275/A5VQ-7Q66

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research, and Evaluation*, *7*, 10. https://doi.org/10.7275/q7rm-gg74

Mutton, T., Hagger, H., & Burn, K. (2011). Learning to plan, planning to learn: The developing expertise of beginning teachers. *Teachers and Teaching*, *17*(4), 399–416. https://doi.org/10.1080/13540602.2011.580516

Ndihokubwayo, K., Byukusenge, C., Byusa, E., Habiyaremye, H. T., Mbonyiryivuze, A., & Mukagihana, J. (2022). Lesson plan analysis protocol (LPAP): A useful tool for researchers and educational evaluators. *Heliyon*, *8*(1), e08730. https://doi.org/10.1016/j.heliyon.2022.e08730

Neumann, K., Härtig, H., Harms, U., & Parchmann, I. (2017). Science teacher preparation in Germany. In J. E. Pedersen, T. Isozaki, & T. Hirano (Eds.), *Model science teacher preparation programs: An international comparison of what works* (pp. 29–52). Information Age Publishing Inc.

Newton, P. E. (2013). Two kinds of argument? *Journal of Educational Measurement*, *50*(1), 105–109.

Otero, V. K., & Nathan, M. J. (2008). Preservice elementary teachers' views of their students' prior knowledge of science. *Journal of Research in Science Teaching*, *45*(4), 497–523. https://doi.org/10.1002/tea.20229

Ozogul, G., Olina, Z., & Sullivan, H. (2008). Teacher, self and peer evaluation of lesson plans written by preservice teachers. *Educational Technology Research and Development*, *56*(2), 181–201. https://doi.org/10.1007/s11423-006-9012-7

Panadero, E., Andrade, H., & Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: a roadmap of where we are, how we got here, and where we are going. *The Australian Educational Researcher*, *45*(1), 13–31. https://doi.org/10.1007/s13384-018-0258-y

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review*, *30*, 100329. https://doi.org/10.1016/j.edurev.2020.100329

Park, S., Jang, J. Y., Chen, Y. C., & Jung, J. (2011). Is pedagogical content knowledge (PCK) necessary for reformed science teaching? Evidence from an empirical study. *Research in Science Education*, *41*(2), 245–260. https://doi.org/10.1007/s11165-009-9163-8

Park, S., & Oliver, J. S. (2008). National Board Certification (NBC) as a catalyst for teachers' learning about teaching: The effects of the NBC process on candidate teachers' PCK development. *Journal of Research in Science Teaching*, *45*(7), 812–834. https://doi.org/10.1002/tea.20234

Park, S., & Suh, J. K. (2019). The PCK Map Approach to capturing the complexity of enacted PCK (ePCK) and pedagogical reasoning in science teaching. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 185–197). Springer. https://doi.org/10.1007/978-981-13-5898-2_8

Patton, M. Q. (1990). *Qualitative evaluation and research methods [Nachdr.]* (2nd ed.). Sage.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448. https://doi.org/10.1080/02602930902862859

Reynolds, W. M., & Park, S. (2021). Examining the relationship between the educative teacher performance assessment and preservice teachers' pedagogical content knowledge. *Journal of Research in Science Teaching*, *58*, 721–748. https://doi.org/10.1002/tea.21676

Ruys, I., Keer, H. V., & Aelterman, A. (2012). Examining pre-service teacher competence in lesson planning pertaining to collaborative learning. *Journal of Curriculum Studies*, *44*(3), 349–379. https://doi.org/10.1080/00220272.2012.675355

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Sato, M. (2014). What is the underlying conception of teaching of the edTPA. *Journal of Teacher Education*, *65*(5), 421–434. https://doi.org/10.1177/0022487114542518

Scherer, R. (2017). The quest for the holy grail of validity in science assessments: A comment on Kampa and Köller (2016) "German National Proficiency Scales in Biology: Internal Structure, Relations to General Cognitive Abilities and Verbal Skills". *Science Education*, *101*(5), 845–853. https://doi.org/10.1002/sce.21278

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competence rubric. *Communication Education*, *61*(3), 205–233. https://doi.org/10.1080/03634523.2012.670709

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

SenBJF. (2017). Handbuch Vorbereitungsdienst: Materialien für den reformierten Berliner Vorbereitungsdienst [Guide for teacher induction in Berlin(6th ed.). https://www.berlin.de/sen/bildung/fachkraefte/lehrerausbildung/vorbereitungsdienst/handbuch_vorbereitungsdienst.pdf

Settlage, J. (2013). On acknowledging PCK's shortcomings. *Journal of Science Teacher Education*, *24*(1), 1–12. https://doi.org/10.1007/s10972-012-9332-x

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2019). Standards für die Lehrerbildung: Bildungswissenschaften [Standards for Teacher Education]. Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 16.05.2019. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf

Stender, A., Brückmann, M., & Neumann, K. (2017). Transformation of Topic-Specific professional knowledge into personal pedagogical content knowledge through lesson planning. *International Journal of Science Education*, *39*(12), 1690–1714. https://doi.org/10.1080/09500693.2017.1351645

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. *Frontiers in Education*, *4*, 89. https://doi.org/10.3389/feduc.2019.00089

Turley, E. D., & Gallagher, C. W. (2008). On the "Uses" of rubrics: Reframing the Great Rubric Debate. *The English Journal*, *97*(4), 87–92.

de Vaus, D. A. (2014). *Surveys in social research*. Social research today (6th ed.). Routledge.

Vogelsang, C., Kulgemeyer, C., & Riese, J. (2022). Learning to plan by learning to Reflect?—Exploring relations between professional knowledge, reflection skills, and planning skills of preservice physics teachers in a one-semester field experience. *Education Sciences*, *12*(7), 479. https://doi.org/10.3390/educsci12070479

Vogelsang, C., & Riese, J. (2017). Wann ist eine Unterrichtsplanung 'gut'?—Planungsperformanz in Praxisratgebern zur Unterrichtsplanung [What makes good lesson planning? Planning performance in advice literature for lesson planning]. In S. Wernke & K. Zierer (Eds.), *Die Unterrichtsplanung: ein in Vergessenheit geratener Kompetenzbereich?! Status Quo und Perspektiven aus Sicht der empirischen Forschung* (pp. 47–61). Verlag Julius Klinkhardt.

Wegener, D. T., & Blankenship, K. L. (2007). Ecological validity. In R. F. Baumeister & K. D. Vohs (Eds.), *Encyclopedia of social psychology* (pp. 275–277). Sage Publications.

Weitzel, H., & Blank, R. (2020). Pedagogical content knowledge in peer dialogues between pre-service biology teachers in the planning of science lessons. Results of an intervention study. *Journal of Science Teacher Education*, *31*(1), 75–93. https://doi.org/10.1080/1046560X.2019.1664874

Westerman, D. A. (1991). Expert and novice teacher decision making. *Journal of Teacher Education*, *42*(4), 292–305.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (expanded 2nd ed.). Association for Supervision and Curriculum Development. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=133964

Wilson, M., de Boeck, P., & Carstensen, C. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects* (pp. 91–120). Hogrefe & Huber.

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

Zaragoza, A., Seidel, T., & Hiebert, J. (2021). Exploring preservice teachers' abilities to connect professional knowledge with lesson planning and observation. *European Journal of Teacher Education*, 1–20. https://doi.org/10.1080/02619768.2021.1996558

Zaragoza, A., Seidel, T., & Santagata, R. (2023). Lesson analysis and plan template: Scaffolding preservice teachers' application of professional knowledge to lesson planning. *Journal of Curriculum Studies*, *55*(2), 13–8152. https://doi.org/10.1080/00220272.2023.2182650

Zhai, X., Krajcik, J., & Pellegrino, J. W. (2021). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology*, *30*(2), 298–312. https://doi.org/10.1007/s10956-020-09879-9

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Großmann, L., & Krüger, D. (2024). Assessing the quality of science teachers' lesson plans: Evaluation and application of a novel instrument. *Science Education*, *108*, 153–189. https://doi.org/10.1002/sce.21832