

# Mathematics teachers' professional noticing: Transfer of a video-based competence assessment instrument into teacher education for evaluation purposes

Jonas Weyers  · Johannes König  · Benjamin Rott  ·  
Gilbert Greefrath  · Katrin Vorhölter  · Gabriele Kaiser 

Received: 28 February 2022 / Revised: 20 December 2022 / Accepted: 7 February 2023 / Published online: 9 June 2023  
© The Author(s) 2023

**Abstract** Teacher noticing has become widely accepted as a principal component of teacher competence; it is supported during university teacher education in many activities. However, only a few high-quality standardized measurement instruments exist that capture noticing and allow valid interpretations of how its development depends on factors within university teacher education. The present study is based on a video-based test instrument that has been developed to enable a standardized study of the noticing of practicing mathematics teachers—that is, their perception, interpretation, and decision-making skills—with respect to subject-specific and general pedagogical issues in secondary mathematics classrooms. This study examines

---

✉ Jonas Weyers · Prof. Dr. Johannes König  
Faculty of Human Sciences, University of Cologne, Gronewaldstraße 2a, 50931 Cologne, Germany  
E-Mail: [jonas.weyers@uni-koeln.de](mailto:jonas.weyers@uni-koeln.de)

Prof. Dr. Johannes König  
E-Mail: [johannes.koenig@uni-koeln.de](mailto:johannes.koenig@uni-koeln.de)

Prof. Dr. Gilbert Greefrath  
Institute of Mathematics Education and Computer Science Education, University of Münster,  
Henriette-Son-Str. 19, 48149 Münster, Germany  
E-Mail: [greefrath@wwu.de](mailto:greefrath@wwu.de)

PD Dr. Katrin Vorhölter  
Faculty of Education, University of Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany  
E-Mail: [Katrin.Vorhoelter@uni-hamburg.de](mailto:Katrin.Vorhoelter@uni-hamburg.de)

Prof. Dr. Gabriele Kaiser  
Faculty of Education, University Hamburg/Nord University Bodø, Von-Melle-Park 8, 20146 Hamburg,  
Germany  
E-Mail: [gabriele.kaiser@uni-hamburg.de](mailto:gabriele.kaiser@uni-hamburg.de)

Prof. Dr. Benjamin Rott  
Faculty of Mathematics and Natural Sciences, University of Cologne,  
Gronewaldstraße 2, 50931 Cologne, Germany  
E-Mail: [Benjamin.rott@uni-koeln.de](mailto:Benjamin.rott@uni-koeln.de)

how this instrument developed for in-service teachers can be used for pre-service teachers at the master's degree level. Based on a sample of 313 pre-service mathematics teachers enrolled in six German universities, the study investigates (1) the instrument's internal structure (scaling models based on item response theory) and (2) its association with relevant factors within university teacher education. The results reveal that a scaling model based on the three noticing facets (perception, interpretation, and decision-making) was superior to a one-dimensional scaling model. Opportunity to learn in mathematics education and average grades in final secondary school examinations were shown to be significant predictors of test performance. However, there was no effect for university-specific opportunity to learn in general pedagogy or pedagogical experiences outside teacher education. Overall, the results suggest that the measurement instrument can be used to assess pre-service mathematics teachers' noticing in university teacher education.

**Keywords** Teacher noticing · Teacher professional vision · Test · Teacher expertise · Teacher education

### **Professionelle Unterrichtswahrnehmung von Mathematiklehrkräften – Zum Einsatz eines videobasierten Testinstruments in der Lehramtsausbildung zu Evaluationszwecken**

**Zusammenfassung** Die professionelle Unterrichtswahrnehmung (konzeptualisiert als Noticing) gilt zunehmend als zentrale Komponente der Lehrkräftekompetenz, die im Rahmen zahlreicher Aktivitäten während der universitären Lehrkräfteausbildung gefördert wird. Allerdings liegen nur wenige qualitativ hochwertige Testinstrumente vor, die die Unterrichtswahrnehmung erfassen und valide Rückschlüsse auf deren Entwicklung in Abhängigkeit universitärer Ausbildungsfaktoren erlauben. Der vorliegende Beitrag basiert auf einem videobasierten Testinstrument, das ursprünglich für praktizierende Mathematiklehrkräfte entwickelt wurde, um die professionelle Unterrichtswahrnehmung, d.h. Wahrnehmung, Interpretation und Entscheidungsfindung, mit Schwerpunkt auf mathematikspezifische sowie allgemeinpädagogische Aspekte des Mathematikunterrichts in der Sekundarstufe zu erfassen. Untersucht wurde, inwieweit dieses Instrument auch für Lehramtsstudierende in der Masterphase einsetzbar ist. Anhand einer Stichprobe von 313 angehenden Mathematiklehrkräften an sechs deutschen Universitäten prüft der vorliegende Beitrag (1) die faktorielle Struktur des Instruments (mithilfe von Item-Response-Theorie Skalierungsmodellen) sowie (2) seine Verbindung zu relevanten Einflussgrößen in der Lehrkräfteausbildung. Die Ergebnisse zeigen, dass das dreidimensionale Skalierungsmodell basierend auf den drei Noticing-Facetten (Wahrnehmung, Interpretation und Entscheidungsfindung) einem eindimensionalen Skalierungsmodell in dieser Stichprobe überlegen ist. Mathematikdidaktische Lerngelegenheiten und Abiturnote konnten als signifikante Prädiktoren der Testleistung identifiziert werden. Hingegen war kein Effekt durch universitäre erziehungswissenschaftliche Lerngelegenheiten sowie durch pädagogische Vorerfahrung außerhalb der Lehrkräfteausbildung nachweisbar. Insgesamt legen die Ergebnisse nahe, dass das Instrument für Studierende

in der Mathematiklehrkräfteausbildung zur Evaluation von professioneller Unterrichtswahrnehmung geeignet ist.

**Schlüsselwörter** Noticing · Professionelle Unterrichtswahrnehmung · Testinstrument · Lehrkräfteexpertise · Lehrer\*innenbildung

## 1 Introduction

Over the last two decades, research on teacher competence has focused primarily on the acquisition of knowledge, understood as a prerequisite for successful teaching and measured with standardized knowledge tests (e.g., Baumert and Kunter 2013; Kunina-Habenicht et al. 2013; Voss et al. 2015). However, reforms in teacher education in many parts of the world, including Germany, have included a stronger orientation toward professional practice, as shown in the implementation of extensive practical learning opportunities in schools (Ulrich and Gröschner 2020) and in the standards specified by the Standing Conference of the Ministers of Education and Cultural Affairs (KMK), which recommended that theoretical concepts be illustrated through the use of practical examples, simulations of teaching situations, and analysis of videotaped instructional practice (KMK 2019). Consequently, to evaluate the effectiveness of university teacher education, test instruments are needed that not only assess knowledge but also the application of knowledge into practice.

One promising measure of pre-service teachers' competence as an outcome of teacher education is the contextualized assessment of competence, which embeds test items into a practical context, commonly using videotaped instructional practice (Gold and Holodynski 2017; Seidel and Stürmer 2014; Wiens et al. 2013). Contextualized assessment aims at providing a measure of competence that is related more closely to performance and reflects implicit instead of inert knowledge (Neuweg 2015). A central framework underlying contextualized assessment is *teacher noticing*, broadly defined as “specialized ways in which teachers observe and make sense of classroom events and instructional details” (Choy and Dindyal 2020).<sup>1</sup> In the current discourse, a construct similar to teacher noticing has been established using the term *professional vision*. Following Santagata et al. (2021), professional vision does not necessarily indicate a different theoretical perspective on teacher noticing. For this reason, the terms *noticing* and *professional vision* are used synonymously for the present paper with both representing a set of mental processes that teachers engage in (see also Sect. 2.1).

Especially in the domain of mathematics teaching, noticing has become widely accepted as a component of teachers' professional competence (Jacobs et al. 2010; Santagata et al. 2021; Sherin et al. 2011a). However, the standardized measurement of noticing is challenging, and only a few high-quality test instruments have been developed and implemented for pre-service teachers.

---

<sup>1</sup> Concerning research on noticing and professional vision, there are differences in terminology and quality of processes differentiated. When referring to specific studies, we therefore use the term which is proposed by the authors.

Against this background, we draw on a standardized noticing test instrument developed in a German follow-up study to the international comparative study Teacher Education and Development Study in Mathematics (TEDS-M), the Teacher Education and Development Study in Mathematics Follow-Up (TEDS-FU; Blömeke et al. 2014). The TEDS-FU Video Test captures secondary mathematics teachers' noticing, which is conceptualized as perception, interpretation, and decision-making skills. While the instrument was successfully implemented for in-service teachers within the projects TEDS-Validate and TEDS-Instruct (Kaiser and König 2020), its use for pre-service teachers has not been explored yet.

Our study builds on the concept of *transfer*, which generally describes the process of a phenomenon or construct being conveyed to another context; more specifically for education science, it denotes the dissemination of innovations from research into educational practice (Gräsel 2010). The TEDS-FU Video Test is transferred to a new context, namely initial teacher education, and—concerning the specific understanding of transfer—its use to measure a learning outcome of university teacher education is investigated. For this purpose, specific validity evidence needs to be provided with respect to the particular group of pre-service teachers (American Educational Research Association [AERA] et al. 2014). This procedure is crucial, since pre-service teachers who have received little explicit training in teaching are not necessarily capable of analyzing video-taped instruction by connecting theoretical concepts and pedagogical practice (the “theory-practice-gap”; see Korthagen 2010).

## 2 Theoretical background

### 2.1 Teacher noticing as part of professional competence

Research on teacher noticing is framed by heterogeneous conceptualizations and terminologies. In a systematic literature review, Santagata et al. (2021) identified four perspectives on teacher noticing: (1) a *cognitive-psychological perspective* that conceptualizes noticing as a set of mental processes that teachers engage in during instruction (e.g., van Es and Sherin 2002), (2) a *socio-cultural perspective*, often associated with the term “professional vision,” that points out the role of social interaction within groups of professionals in shaping a common perception and understanding of meaningful events (Goodwin 1994), (3) a *discipline-specific perspective* that conceptualizes noticing as a set of practices teachers engage in to support their own sensitivity (Mason 2002), and (4) an *expertise-related perspective* that highlights differences between novice teachers and experts with respect to their ways of seeing and making sense of observed instructional practice (Berliner 1988). The practice of measuring teacher noticing with standardized test instruments was especially influenced by the cognitive-psychological perspective described in detail below.

Seen from the cognitive-psychological perspective, noticing is conceived of as a set of closely interrelated mental processes, called *noticing facets*, that teachers engage in during instruction and “through which teachers manage the ‘blooming,

buzzing confusion of sensory data' with which they are faced" (Sherin et al. 2011b, p. 7). Three noticing facets were differentiated by van Es and Sherin (2002): "(a) identifying what is important or noteworthy about a classroom situation; (b) making connections between the specifics of classroom interactions and the broader principles of teaching and learning they represent; and (c) using what one knows about the context to reason about classroom interactions" (p. 573). Focusing on noticing children's mathematical thinking, this approach was restructured and expanded by Jacobs et al. (2010), who differentiated the noticing facets as (a) "attending to children's strategies," (b) "interpreting children's mathematical understandings," and (c) "deciding how to respond on the basis of children's understanding" (pp. 172–173). However, no consensus has been reached so far with respect to how many and what kind of facets are relevant to conceptualize and investigate teacher noticing (Dindyal et al. 2021).

Given the heterogeneity of perspectives and conceptualizations of teacher noticing, the development of a consistent theoretical framework serving as basis for test development is challenging. The present study builds upon the widely accepted theoretical framework by Blömeke et al. (2015a) of competence as a continuum, in which a set of situation-specific skills, that is, perception, interpretation, and decision-making, is conceptualized as mediator between dispositions (e.g., knowledge or beliefs) and performance. This model can be seen as extension of cognitive approaches to competence, primarily focusing on professional knowledge (e.g., Baumert and Kunter 2013). Noticing, in our framework, is thus seen as part of professional competence and conceptualized as a set of situations-specific skills, which is comparable to the mental processes focused on within the psychological perspective on teacher noticing (e.g., Jacobs et al. 2010). For our own framework, we use the more neutral term "noticing facet" when referring to the different qualities of skills/processes.

The competence as a continuum model was transferred to mathematics teaching by Kaiser et al. (2015, p. 374) conceptualizing teacher noticing as "(a) *Perceiving* particular events in an instructional setting, (b) *Interpreting* the perceived activities in the instructional setting and (c) *Decision-making*, either as anticipating responses to students' activities or as proposing alternative instructional strategies". Although this model's focus is on mathematics teaching, the scope of noticing is broadened by considering subject-specific as well as generic pedagogical issues in whole lessons, including noticing of students' and teachers' actions. For this framework, the first facet is termed "perception" instead of "attending" with reference to the research on teacher expertise. Within this research strand, perception denotes teachers' processing of relevant sensory information (e.g., Carter et al. 1988). While attending emphasizes the selectivity of information processing (e.g., attending to a relevant detail within a complex perceptual field), the term perception implies a stronger focus on perceiving (and remembering) clearly discernable events.<sup>2</sup> Consequently, the accurate perception of classroom events does not necessarily require professional knowledge (or experience), even though knowledge inevitably shapes perception.

---

<sup>2</sup> The accurate perception of classroom events necessarily encompasses attentional processes. The use of the term "perception" in this framework should, therefore, be understood as stronger focus on perceptual processes compared to attentional processes with both being involved.

The second facet, *interpretation*, refers to the teacher's thinking about what they have observed using their knowledge and experience, thereby "relating observed events to abstract categories and characterizing what they see in terms of familiar instructional episodes" (Sherin et al. 2011b, p. 5). Based on their perceptions and interpretations, teachers must determine appropriate instructional responses to classroom events; this is the third facet, *decision-making*.

## 2.2 Standardized testing of teacher noticing: Test design and validation

Standardized noticing tests commonly include the presentation of classroom artifacts—usually short videos of classroom practice—combined with rating items or open questions to capture the various noticing facets (Jacobs et al. 2010; Kaiser et al. 2015; Star and Strickland 2008). Since noticing does not represent one homogeneous construct, measurement instruments commonly target a specific *focused domain*, which is related to subject-specific aspects, such as children's mathematical thinking (Jacobs et al. 2010) or instructional support in primary science teaching (Todorova et al. 2017), or generic pedagogical aspects (Seidel and Stürmer 2014; Wiens et al. 2020).

Given that this measurement approach is comparably new, the investigation of validity is of particular relevance. Following AERA et al. (2014), validity is understood as a unitary concept, which "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Consequently, researchers are required to specify the intended interpretation(s) of tests scores and the intended test use—including a precise definition of the underlying construct—and collect theoretical and/or empirical validity evidence to support these interpretation(s).

For valid interpretations of noticing tests, it should be investigated whether the theoretical conceptualization of noticing, especially regarding the differentiation of noticing facets, corresponds to the measurement using factor analysis or item response theory (IRT). However, previous findings vary. For example, Seidel and Stürmer (2014) found that a three-dimensional model, distinguishing the facets description<sup>3</sup>, explanation, and prediction, fitted the data better than a one-dimensional model even though the intercorrelations were large ( $0.77 \leq r \leq 0.89$ ). Measuring perception, interpretation, and decision-making, Bastian et al. (2021) favored a three-dimensional over a one-dimensional model with high latent correlations between perception and interpretation (0.814), and interpretation and decision-making (0.815), but a lower correlation between perception and decision-making ( $r=0.462$ ). By contrast, other studies' findings are interpreted in favor of a unidimensional structure (Gold and Holodynski 2017; Meschede et al. 2015). For example, Meschede et al. (2015) report that describing and interpreting are almost inseparable ( $r=0.99$ ).

<sup>3</sup> On the empirical level, the operationalization of "describing" as a facet of professional vision is similar to the operationalization of perceiving or attending within noticing terminology.

## 2.3 Noticing as a learning outcome of teacher education

Using amongst others a noticing test for the evaluation of teacher education programs requires evidence that a substantial proportion of the variance in test scores can be explained by relevant factors in teacher education. In line with the educational concept of learning opportunities in teacher education (Floden 2002; Schmidt et al. 2011), the acquisition of professional competence is conceptualized as an interplay of (1) pre-service teachers' individual prerequisites and (2) their perception of having been exposed to formal opportunity to learn (OTL). Considering both aspects, the following subsections give reason for the variables selected for the present validation study and summarize existing evidence.

### 2.3.1 Individual prerequisites

The *average grade in the final secondary school examinations* is commonly used as a distal indicator of cognitive abilities in research on teacher competence (e.g., Kunina-Habenicht et al. 2013). However, the average grade is further related to knowledge, (academic) motivation and learning strategies (Mayr 2010) and predictive for future academic achievement (Trapmann et al. 2007).

As a broad indicator of academic ability, the average grade in final secondary school examinations has also been shown to predict teachers' professional knowledge regarding general pedagogy (Kunina-Habenicht et al. 2013) as well as teachers' subject-specific content knowledge and pedagogical content knowledge for several domains (König et al. 2018; Lindl and Krauss 2017). As noticing has been conceptualized as knowledge-based, the average grade should therefore also be related to the acquisition of noticing. However, existing findings vary: An effect of the average grade on noticing was found by Wiens and Gromlich (2018) ( $\beta = 0.17$ ) as well as by Todorova et al. (2017) ( $\beta = -0.25/-0.31$ ; lower grades indicate better performance). By contrast, other studies did not find such effects (Stürmer et al. 2015; Wiens et al. 2013), suggesting that the relationship depends on the investigated sample and the specific operationalization of noticing.

Before and during their studies, pre-service teachers can gain *pedagogical experience* in such contexts as private tutoring and coaching sports teams. These activities can be conceptualized both as informal learning opportunities and as individual prerequisites that facilitate the acquisition of professional knowledge (König et al. 2012; Kunina-Habenicht et al. 2013). Pedagogical experience in the context of teaching might also promote pre-service teachers' noticing by providing opportunities for using the acquired knowledge in situations of pedagogical action.

However, some cross-sectional studies have not found a correlation between noticing and pedagogical experience, internship experience, or teaching experience (Jamil et al. 2015; Stürmer et al. 2015; Todorova et al. 2017); although, teaching practice can generally support pre-service teachers' noticing, which has been shown for long-term teaching internships (e.g., Mertens and Gräsel 2018,  $d = 0.79$ ).

### 2.3.2 Use of opportunity to learn

OTL can be broadly defined as experiences that aim to achieve a learning outcome (Tatto et al. 2008). There is substantial variation in the perceived amount of OTL for mathematics pedagogy and teaching mathematics among pre-service teachers (Christiansen and Erixon 2021). However, only a few studies have investigated the influence of program features within teacher education on the acquisition of noticing (Stürmer et al. 2015; Todorova et al. 2017; Wiens et al. 2013). For example, Stürmer et al. (2015) showed that noticing conceptualized as professional vision was associated with the number of generic pedagogical courses ( $\beta = 0.31$ ). Furthermore, Todorova et al. (2017) found that pre-service teachers with a study focus on science teaching outperformed their colleagues with respect to noticing science-specific aspects ( $\beta = 0.30/0.33$ ).

In line with international studies on school achievement and teacher competence, OTL can be operationalized by the specific content a learner has dealt with up to a certain time (Kunina-Habenicht et al. 2013; Schmidt et al. 2011). The amount of OTL experienced is related to the acquisition of professional knowledge during teacher education; thus, OTL is an appropriate variable to use for validation purposes with respect to measures of learning outcomes (König et al. 2018; Kunina-Habenicht et al. 2013). As differentiated assessments of OTL have not been linked to teacher noticing until now, this makes them an interesting measure for validation purposes and for exploring the effects of teacher education on noticing.

## 3 Research questions and background of the study

In recent years, considerable efforts have been made to develop video-based instruments that enable the contextualized assessment of teachers' competence (e.g., Gold and Holodyski 2017; Jamil et al. 2015; Seidel and Stürmer 2014). Our study draws on an instrument developed within the study TEDS-FU, namely the TEDS-FU Video Test, which targeted early career secondary mathematics teachers' noticing skills.

Within the study TEDS-M, teachers' competence was addressed using standardized knowledge tests. To evaluate teachers' competence more closely connected to teaching practice, the conceptual framework of TEDS-M was extended within TEDS-FU by considering teachers' situation-specific skills—that is, perception, interpretation, and decision-making skills—and assessing them using video-based test instruments (Kaiser et al. 2015). In TEDS-FU, the original participants, who had been at the end of their teacher education when participating in TEDS-M, were approached another time after 2.5–3 years of work as early-career teachers. The test development was accompanied by curricular analyses to ensure the accuracy of the mathematical content and expert workshops to discuss the suitability of the test items and instructional events presented in the videos (Kaiser et al. 2015; Hoth et al. 2016).

Test performance was empirically correlated with professional knowledge (Blömeke et al. 2015b). In further studies, namely TEDS-Instruct and TEDS-Validate, the TEDS-FU Video Test was used with practicing teachers with different



lengths of teaching experience (Bastian et al. 2021). These findings suggest that the TEDS-FU Video Test can be validly interpreted as a measure of in-service teachers' noticing skills. However, it remains an open question whether the instrument can be used with pre-service teachers who have little teaching experience.

Our study aims to provide specific validity evidence that the TEDS-FU Video Test can be used with pre-service teachers and measures their noticing skills as one learning outcome of teacher education. For this purpose, we focus (1) on validity evidence based on the internal test structure and (2) associations with relevant factors within university teacher education.

**RQ1** Does the TEDS-FU Video Test reliably measure pre-service teachers' noticing skills—perception, interpretation, and decision-making skills—as the three interrelated facets of noticing?

Evidence based on internal test structure is crucial to the certainty that a reliable measure of the differentiated facets is provided for pre-service teachers, and measurement is not affected by limited variance. Therefore, a one-dimensional scaling model—noticing as one holistic facet—is compared to a three-dimensional model that distinguishes pre-service teachers' perception, interpretation, and decision-making skills. We hypothesize that the three-dimensional model is superior to the one-dimensional model.

To provide validity evidence, the pattern of intercorrelations should correspond to theoretical presumptions (see AERA et al. 2014). Perception and interpretation are discussed as closely related (Sherin et al. 2011b), both being informed by professional knowledge (e.g., Wolff et al. 2021). We thus predict a high correlation between these noticing facets. Similarly, since teachers' decision-making should be based on a sound interpretation (Bastian et al. 2021; Jacobs et al. 2010), we also expect a high correlation between interpretation and decision-making. By contrast, pre-service teachers may perceive and remember discernable features of the classroom without being able to propose an adequate response. So, we expect only a moderate correlation between perception and decision-making.

**RQ2** Can pre-service teachers' noticing scores be explained by (a) the participants' individual prerequisites, namely, the average grade in the final secondary school examinations and pedagogical experience, or (b) the participants' use of formal OTL?

We expect pre-service teachers with higher academic ability—indicated by the average grade—to score higher in noticing, as their better prerequisites support them to acquire and apply knowledge. We further expect teaching experiences (e.g., private tutoring) but not nurturing experiences (e.g., caring for younger brothers and sisters) to correlate with noticing, as only teaching experiences provide opportunities to use the theoretical knowledge acquired for reflecting on teaching situations. As formal OTL provide situations for acquiring and possibly applying professional knowledge, we expected that OTL in general pedagogy and mathematics education predicts noticing. With reference to previous findings, we expect small effect sizes for all factors considered. However, the explained variance should be taken into account for the evaluation of validity evidence.

**Table 1** Demographic statistics

| University   | <i>n</i> | Gender<br>(female; %) | Age<br><i>M</i> ( <i>SD</i> ) | Study<br>semester | Average<br>grade | Teacher education program (%) |      |     |     |
|--------------|----------|-----------------------|-------------------------------|-------------------|------------------|-------------------------------|------|-----|-----|
|              |          |                       |                               |                   |                  | I                             | II   | III | IV  |
| Hamburg      | 62       | 68                    | 27 (4)                        | 9.2 (3.39)        | 1.8 (0.37)       | 66                            | 7    | 27  | 0   |
| Würzburg     | 57       | 56                    | 22 (2)                        | 6.3 (1.40)        | 2.1 (0.64)       | 67                            | 33   | 0   | 0   |
| Vechta       | 33       | 61                    | 24 (2)                        | 7.8 (1.67)        | 2.7 (0.37)       | 100                           | 0    | 0   | 0   |
| Cologne      | 79       | 54                    | 25 (3)                        | 9.5 (2.56)        | 1.9 (0.58)       | 19                            | 78.5 | 0   | 2.5 |
| Paderborn    | 16       | 81                    | 25 (3)                        | 8.8 (2.11)        | 2.2 (0.69)       | 50                            | 50   | 0   | 0   |
| Münster      | 66       | 56                    | 25 (5)                        | 9.8 (2.41)        | 2.1 (0.56)       | 26                            | 68   | 0   | 6   |
| <i>Total</i> | 313      | 60                    | 25 (4)                        | 8.6 (2.69)        | 2.1 (0.61)       | 49                            | 44   | 5   | 2   |

*Average grade* average grade in final secondary school examinations (lower values indicate better performance), *I* lower secondary school, *II* lower and upper secondary school, *III* special needs education, *IV* vocational school

## 4 Methodology

### 4.1 Sample

A sample of 313 pre-service mathematics teachers was surveyed between spring 2019 and fall 2020 at six German universities. Table 1 shows the demographic statistics of the present sample. Pre-service teachers were recruited before they entered their first long-term school internship and so had little teaching experience in the context of university teacher education. For all universities except Würzburg, the internships took place during the master's degree phase. At the University of Würzburg, the study program is organized as a state examination and not divided into bachelor's and master's degrees. Therefore, participants from this university were recruited before entering their study-related teaching internships in the fourth to sixth study semesters.

The participants were contacted by the lecturers of their courses, which focused on mathematics teaching preparatory for the long-term teaching internships. They received an internet link via e-mail that led to an online platform hosting the survey, including noticing tests and questions on supplemental information. Completing the questionnaire took approximately 90 min, and participants were reimbursed with a financial compensation of 15 Euros. Data collection and processing was in accordance with the requirements of the General Data Protection Regulation.

### 4.2 Measures

**Teacher noticing** Pre-service teachers' noticing skills were assessed using the TEDS-FU Video Test (Kaiser et al. 2015), which includes three scripted video vignettes about 3.5 min long: (1) *Frog King* based on a German fairy tale, (2) *Box*, and (3) *Solids*. Scripted vignettes were used, rather than videos of authentic instruction, to ensure a sufficient density of mathematics and generally pedagogically relevant events. The vignettes show compilations of ninth-grade mathematics lessons in different school types that cover a wide range of mathematical topics (e.g., volume calculations, functions, and surfaces) and different instructional phases. Before

**Table 2** Item number and distribution of items across noticing facets

| <i>Noticing facet</i> | <i>nRating Items</i> | <i>nOpen response items</i> | <i>nItems (total)</i> |
|-----------------------|----------------------|-----------------------------|-----------------------|
| Perception            | 19                   | 5                           | 24                    |
| Interpretation        | 22                   | 20                          | 42                    |
| Decision-making       | –                    | 11                          | 11                    |
| <i>Total</i>          | 41                   | 36                          | 77                    |


watching each vignette, the participants received some information about the students, the learning context, and the mathematical topics. Participants were permitted to watch each vignette only once.

Since a detailed description of the three video vignettes can be found in previous publications (see Kaiser et al. 2015), we restrict ourselves to describing one vignette only. *Box* refers to a secondary mathematics classroom of academic-track ninth-grade students who are asked to compute the volume of an open box made from a rectangular sheet with four congruent squares cut off the corners. The volume of the box can be determined based on a function of the size of the cut-off squares. Three pairs of students are shown solving the task in diverse ways. The results are then collected in the whole-class discussion.

After each vignette, rating items and open response items were administered to access the participants' perception, interpretation, and decision-making skills (see Table 2) focusing on both subject-specific and general pedagogical aspects of mathematics teaching. For rating items, the participants indicated the extent to which they (dis)agreed with statements on the observed practice on four-point Likert scales (*fully correct to not correct at all*).

The items with a focus on perception mainly consisted of rating scales including descriptive statements (e.g., "Most students take an active part in the lesson"). Working on these items required the participants to carefully watch the video clips, but not to draw on their professional knowledge. By contrast, items focusing on interpretation, which included both item formats, required the participants to link the observed practice to broader principles of teaching and learning. For the example item in Fig. 1, the participants had to connect the approaches of three pairs of students shown in the video to different modes of representation (enactive, iconic, and symbolic). Working on such items also requires a certain degree of perceptual processing. However, the items were constructed to explicitly focus on interpretative processes (e.g., by addressing contrasting descriptions or the application of concepts), and the participants' perception was supported using pictures of the teaching situations and short introducing texts. Items focusing on decision-making solely comprised open response items and required the participants to propose possible continuations to the instructional practice observed or to create alternatives to the teacher's actions in the video (see Fig. 2). To create unambiguous items with a clear focus on decision-making, the item texts suggested an interpretation of the relevant situation in the video (e.g., the specification of a learning goal for the class).

The scoring procedure was based on an expert survey. For the rating scales, the participants' answers were coded as "correct" if they matched the expert master rating. Scoring of open response items was conducted with an extensive coding




1                      2                      3

In the video-vignette the working processes of three cooperating pairs have been observed more closely. These working processes are to be examined from two perspectives: (a) mathematics education and (b) pedagogics.

(a) mathematics education perspective  
In each of the three approaches the task is represented and solved mathematically in a specific way.  
Please describe (in note form) the essential aspects of the approaches in a contrasting mode from a mathematics education view.  
Please name – if possible – the corresponding technical terms.

(b) pedagogics perspective  
Please describe (in note form) for each of the three pairs in a contrasting mode the essential aspects of the way the two students cooperated in their work.

**Fig. 1** Open response item targeting interpretation with a focus on mathematics teaching (a) and general pedagogy (b)



The educational standards issued by the German state include process standards referring to mathematical competencies.  
Assume that you have to continue the lesson you have just seen in the video by a sensible task. This task should be well connected to the lesson.  
Please phrase a task you would give to the class...

... if you want to strengthen the students' competencies using mathematical representations.

... if you want to strengthen the students' competencies concerning mathematical reasoning.

**Fig. 2** Open response item focusing on decision-making with respect to mathematics-related aspects of teaching

manual based on the experts' solutions; it resulted in good interrater reliability ( $\kappa_{\text{mean}} = 0.80$ ;  $\kappa_{\text{min}} = 0.47$ ;  $\kappa_{\text{max}} = 1.0$ ).

**Individual prerequisites** In addition to the average grade in the final secondary school examinations (minimum: 4.0; maximum: 1.0), the participants' pedagogical experience was assessed using a measure by König et al. (2013). On five dichotomous items (yes/no), the participants indicated whether they had had a specific pedagogical experience or not. The items aim at nurturing on the one hand and at teaching experience outside of formal studies on the other. Example items and descriptive statistics are shown in Table 3.

**Opportunity to learn** Pre-service teachers' formal OTL was assessed with respect to (1) teaching mathematics (Doll et al. 2018) and (2) general pedagogy (König et al. 2017). The participants had to indicate (yes/no) whether specific content had been treated within their previous teacher training. The content represented central topics of German teacher training within the two areas focused on. Subscales, example items, and descriptive statistics can be found in Table 3. Internal consistency was at least acceptable for all subscales.

**Table 3** Overview of measures for OTL and pedagogical experience

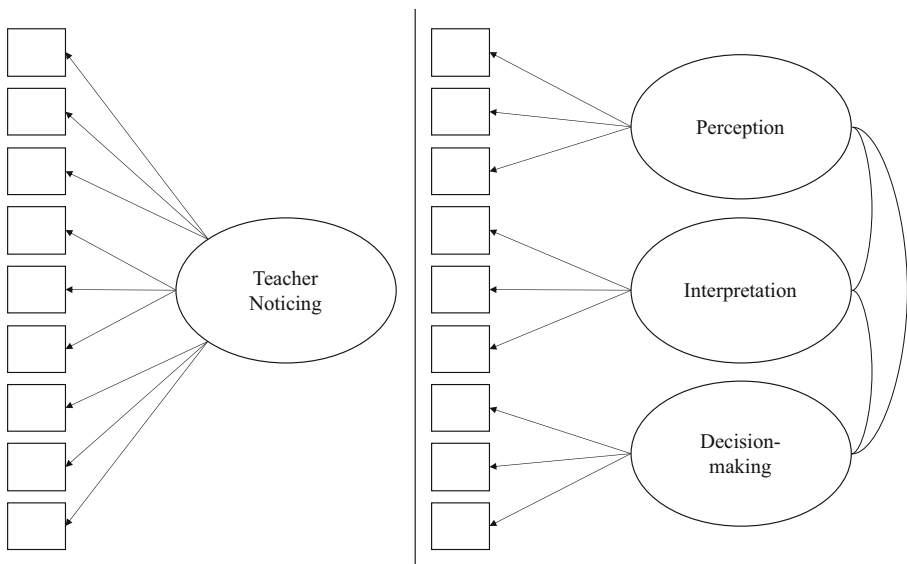
| Scale   | Sample item  | $N_{\text{items}}$ | $M (SD)$    | $\alpha$ |
|---|--|--------------------|-------------|----------|
| <i>Opportunity to learn—General Pedagogy</i>      |  |                    |             |          |
| Adaptivity  | Individual instructional support   | 11                 | 0.65 (0.24) | 0.75     |
| Structuring                                       | Teaching methods   | 9                  | 0.78 (0.24) | 0.76     |
| Classroom management/<br>Motivation               | Classroom rules  | 8                  | 0.52 (0.29) | 0.74     |
| Assessment  | Diagnostics of learning processes  | 9                  | 0.61 (0.34) | 0.87     |
| Total (OTL Pedagogy)                              |  | 37                 | 0.64 (0.21) | 0.90     |
| <i>Opportunity to learn—Mathematics education</i> |  |                    |             |          |
| Basics  | Didactics of algebra   | 20                 | 0.57 (0.19) | 0.76     |
| Adaptivity  | Language sensitive mathematics teaching  | 8                  | 0.33 (0.22) | 0.60     |
| Digital media                                     | App-based learning in mathematics  | 6                  | 0.46 (0.29) | 0.73     |
| Methodology                                       | Problem-oriented teaching in mathematics   | 8                  | 0.50 (0.24) | 0.67     |
| Curricular aspects/<br>Assessment                 | Curricula for mathematics teaching   | 8                  | 0.49 (0.22) | 0.62     |
| Research in teaching mathematics                  | TEDS studies or COACTIV study  | 11                 | 0.31 (0.20) | 0.66     |
| Total (OTL Mathematics)                           |  | 61                 | 0.46 (0.16) | 0.89     |
| <i>Pedagogical experience</i>                     |  |                    |             |          |
| Nurturing   | Caring for children (e.g., brothers and sisters, own children, babysitting, au pair) | 2                  | 0.67 (0.36) | –        |
| Teaching  | Tutoring/Homework supervision—as one-to-one lessons                                  | 3                  | 0.50 (0.29) | –        |

Means represent the relative frequency of content or experiences reported by participants

### 4.3 Data analysis

Test data were scaled based on item response theory (IRT) with *ConQuest* software (Wu et al. 1997) using Rasch models. To investigate the internal test structure of the TEDS-FU Video Test (RQ1), an IRT scaling model was initially estimated with one latent variable. Then, a multidimensional IRT model with three latent variables—(1) perception, (2) interpretation, and (3) decision-making—was specified. The two scaling models (see Fig. 3) were compared with respect to the expected a posteriori/plausible values (EAP/PV) reliability, the weighted likelihood estimates (WLE) reliability, and the theta variance, model deviance, and sample-size-adjusted Bayesian information criterion (BIC).

To explore the relationship between test performance (WLE estimates of person ability), and factors within teacher education (RQ2), multiple regression models were conducted using *Mplus* (Muthén and Muthén 1998–2006). The stratified structure of the sample was considered by using the option “type=complex” and specifying a combined variable of university and teacher education program as a stratum.<sup>4</sup>



**Fig. 3** Scaling models of teacher noticing

<sup>4</sup> Participants were included when data were available for at least 50% of the test items. For the final sample, the proportion of missing values in the test data was small (5%) and person parameters could be estimated based on the available data. Regarding the predictor variables, only few data points were missing (e.g., OTL in mathematics education: 0.1%). For OTL in general pedagogy and the study semester, 12% of the data were missing, since for organizational reasons for one part of the sample this questionnaire was not administered. Cases with missing values on predictor variables were removed from the analysis, so the regression models are based on a sample of around 275 participants.

**Table 4** Comparison of the one-dimensional and the three-dimensional scaling model

| Model                       | BIC       | Deviance  | Para me-<br>ters | Likelihood-ratio test   | WLE reliability                     | EAP/PV reliability                  | Variance                            |
|-----------------------------|-----------|-----------|------------------|---|-------------------------------------|-------------------------------------|-------------------------------------|
| 1-Dim.                      | 25,747.64 | 25,287.95 | 80               | –   | 0.857                               | 0.827                               | 0.424                               |
| P-I-D<br>(3-Dim.)           | 25,732.26 | 25,243.84 | 85               | <b>P-I-D vs. I-Dim.</b><br>$\Delta$ Deviance = 44.11<br>$df=5$<br>$p < 0.001$   | Per 0.669<br>Int 0.800<br>Dec 0.469 | Per 0.749<br>Int 0.868<br>Dec 0.662 | Per 0.454<br>Int 0.453<br>Dec 0.650 |
| <i>Exploratory analysis</i> |           |           |                  |   |                                     |                                     |                                     |
| PI-D<br>(2-Dim.)            | 25,749.88 | 25,278.68 | 82               | <b>PI-D vs. I-Dim.</b><br>$\Delta$ Deviance = 9.27<br>$df=2$<br>$p < 0.01$<br><b>PI-D vs. P-I-D</b><br>$\Delta$ Deviance = 34.84<br>$df=3$<br>$p < 0.001$   | Per/Int 0.840<br>Dec 0.448          | Per/Int 0.850<br>Dec 0.647          | Per/Int 0.441<br>Dec 0.608          |
| P-ID<br>(2-Dim.)            | 25,697.29 | 25,226.10 | 82               | <b>P-ID vs. I-Dim.</b><br>$\Delta$ Deviance = 61.84<br>$df=2$<br>$p < 0.001$<br><b>P-ID vs. P-I-D</b><br>$\Delta$ Deviance = 17.73<br>$df=3$<br>$p < 0.001$ | Per 0.681<br>Int/Dec 0.830          | Per 0.714<br>Int/Dec 0.811          | Per 0.499<br>Int/Dec 0.477          |

For the Likelihood-ratio test, the **bold** model shows better fit  
*BIC* Bayesian Information Criterion, *Deviance*  $-2 \log$  (likelihood ratio),  $\Delta$ Deviance  $\chi^2$ -distributed test statistic, *I-Dim.* noticing as one latent variable (*PI-D*), *P-I-D* perception (*Per*) vs. interpretation (*Int*) vs. decision-making (*Dec*), *PI-D* *Per*/*Int* vs. *Dec*, *P-ID* *Per* vs. *Int*/*Dec*

## 5 Results

### 5.1 RQ1: Internal test structure and reliability

Item analysis was conducted for the one-dimensional and three-dimensional partial credit models. Seven items were removed from analysis since they exceeded a *weighted mean square (WMSQ)* of 1.25 or showed poor item discrimination ( $<0.15$ ). Two further items with critical fit statistics were kept for theoretical reasons. For the remaining items, item discrimination was, on average, good ( $M=0.30$ ,  $\text{min.}=0.15$ ,  $\text{max.}=0.54$ ) and *WMSQs* were in an appropriate range ( $0.88 < WMSQ < 1.13$ ; Bond and Fox 2015).

The results of both scaling models are depicted in Table 4. The model deviance and the corresponding likelihood ratio test revealed that the three-dimensional model fitted the data significantly better than the one-dimensional model. The WLE and EAP/PV reliability can be considered as very good for the one-dimensional model. Regarding the three-dimensional model, the reliability was still acceptable or good for perception and interpretation, but the WLE reliability for decision-making was very low, which is in part due to the smaller number of items used for this dimension (see Table 2).

With respect to the latent intercorrelations, correlation was high between perception and interpretation ( $r_{PI}=0.704$ ) and between interpretation and decision-making ( $r_{ID}=0.730$ ); it was lower between perception and decision-making ( $r_{PD}=0.292$ ). This latter correlation ( $r_{PD}$ ) was significantly<sup>5</sup> lower than  $r_{PI}$  ( $z=-11.928$ ;  $p<0.001$ ) and  $r_{ID}$  ( $z=-12.361$ ;  $p<0.001$ ), which is in line with our hypotheses. In sum, scaling analysis and intercorrelations support the superiority of the three-dimensional model.

In exploratory analyses, two further models were tested with two dimensions respectively: (1) perception and interpretation vs. decision-making (PI-D;  $r_{PI-D}=0.681$ ), (2) perception vs. interpretation and decision-making (P-ID;  $r_{P-ID}=0.671$ ). Both models showed better model fit than the one-dimensional model, and the P-ID even fitted better than the three-dimensional model (see Table 4). This result suggests that combining interpretation and decision-making may provide a more efficient approach of measuring noticing when using this instrument. However, the lower deviance of the P-ID model is partly explained by the low reliability—and the low number of items—focusing on decision-making. To account for possible differences between interpretation and decision-making regarding their relationship with other variables, the three-dimensional model was used for the subsequent analyses.

It should be noted that the items measuring perception mainly have a focus on general pedagogy, while decision-making items predominantly address mathematics teaching. Consequently, the correlation between decision-making and perception may be underestimated. However, further analyses revealed that the low correlation between perception and decision-making is likely not a result of different domains focused on by the items (see Online Resource 1).

<sup>5</sup> Comparisons of correlation coefficients were conducted following Meng et al. (1992).



## 5.2 RQ2: Effects of individual prerequisites and opportunity to learn

The (manifest) correlations between noticing facets and factors within teacher education including individual prerequisites and OTL can be found in Online Resource 2. Study semester, average grade in the final secondary school examinations, and a dichotomous indicator of the teacher education program (0=lower and upper secondary school/vocational school, 1=lower secondary school/special needs education) were included as control variables for all models. Other demographic variables were not included since they did not correlate with noticing. For each facet, four regression models were specified: One model focusing on individual prerequisites, two separate models including OTL in mathematics education or general pedagogy, and one model with all predictors being included. Separate models were specified for the two domains of OTL to avoid a loss of statistical power owing to the high correlation between mathematics teaching and general pedagogy ( $r=0.48$ ).

The results of the multiple regression analysis can be seen in Table 5. Against our expectations, for the perception facet, only a small proportion of the variance was explained. Only the average grade in the final secondary school examinations showed a significant but small effect on perception, with better test performance being associated with a better average grade. Interpretation and decision-making were also significantly predicted by the average grade in the final secondary school examinations. However, against our assumptions, the only effect that could be found for OTL was a small effect of OTL in mathematics education on decision-making. Another very small effect of OTL in mathematics education on interpretation was not significant when controlling for OTL in general pedagogy. For all facets, no effect of pedagogical experience or OTL in general pedagogy was found.

Using a combination of university and education program as a cluster variable, a considerable proportion of variance was found to be on program level for interpretation (ICC=0.11) and decision-making (ICC=0.09), but not for perception (ICC<0.01). Therefore, additional multilevel regression models were specified to explore the effects of OTL on interpretation and decision-making when distinguishing between individual use of OTL and the influence of the teacher education program (i.e., the context effect). The results can be found in Online Resource 3, Table 1, and are comparable to the results of the regression models reported above, except for a very small effect ( $\beta=0.12$ ) of OTL in general pedagogy predicting interpretation on level 1. However, it should be noted that a substantial proportion of variance regarding interpretation on program level was explained when OTL in mathematics education were included in the model (around 45%). Even though this effect was not significant—the small number of clusters reduces the statistical power—this effect can serve as a starting point for further analyses.

To examine the effect of the OTL subscales, a multiple regression model for each subscale was specified, including study semester, teacher education program, and average grade in the final secondary school examinations as control variables (see Table 6). Using separate models accounts for possibly reduced power caused by moderate correlations between the OTL subscale (see Online Resource 2, Tables 2 and 3). Given the increased alpha risk due to the number of models estimated, the significance criterion was reduced by factor 0.1 ( $p<0.005$ ), which is equivalent to

**Table 5** Multiple regression models for perception, interpretation, and decision-making

| Variable                      | Perception |         |         | Interpretation |          |          | Decision-making |          |          |         |          |          |
|-------------------------------|------------|---------|---------|----------------|----------|----------|-----------------|----------|----------|---------|----------|----------|
|                               | M1         | M2      | M3      | M4             | M5       | M6       | M7              | M8       | M9       | M10     | M11      | M12      |
|                               | $\beta$    | $\beta$ | $\beta$ | $\beta$        | $\beta$  | $\beta$  | $\beta$         | $\beta$  | $\beta$  | $\beta$ | $\beta$  | $\beta$  |
| Semester                      | 0.11       | 0.10    | 0.11    | 0.11           | 0.14**   | 0.13*    | 0.14**          | 0.13*    | 0.06     | 0.04    | 0.06     | 0.05     |
| Program                       | 0.04       | 0.02    | 0.03    | 0.01           | 0.13*    | 0.08     | 0.12*           | 0.07     | 0.06     | -0.06   | 0.01     | -0.07    |
| Average grade                 | -0.20**    | -0.19** | -0.19** | -0.19**        | -0.40*** | -0.39*** | -0.40***        | -0.40*** | -0.27*** | -0.26** | -0.27*** | -0.26*** |
| <i>Pedagogical experience</i> |            |         |         |                |          |          |                 |          |          |         |          |          |
| Nurturing                     | -0.04      | -0.05   | -0.06   | -0.06          | -0.01    | 0.00     | -0.02           | -0.03    | -0.04    | -0.06   | -0.06    | -0.07    |
| Teaching                      | 0.06       | 0.05    | 0.06    | 0.05           | 0.00     | 0.00     | 0.01            | 0.00     | 0.06     | 0.05    | 0.06     | 0.05     |
| <i>OTL</i>                    |            |         |         |                |          |          |                 |          |          |         |          |          |
| Mathematics                   | -          | 0.07    | -       | 0.05           | -        | 0.13*    | -               | 0.13     | -        | 0.20**  | -        | 0.25**   |
| Pedagogy                      | -          | -       | 0.08    | 0.06           | -        | -        | 0.06            | 0.01     | -        | -       | 0.05     | -0.05    |
| R <sup>2</sup>                | 0.05       | 0.06    | 0.06    | 0.06           | 0.15     | 0.17     | 0.16            | 0.17     | 0.08     | 0.12    | 0.09     | 0.13     |

$\beta$  standardized regression coefficients

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

**Table 6** Summary of beta coefficients of OTL subscales

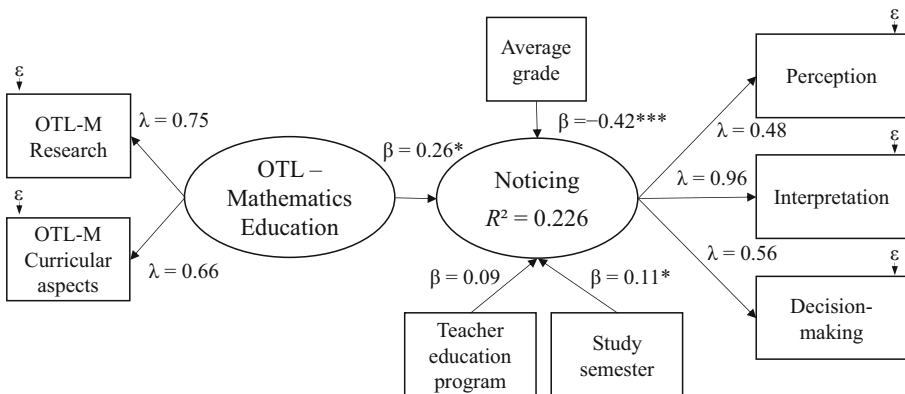
| Domain of OTL         | Subscale                          | Perception |       |       | Interpretation |       |       | Decision-making |        |       |
|-----------------------|-----------------------------------|------------|-------|-------|----------------|-------|-------|-----------------|--------|-------|
|                       |                                   | $\beta$    | $p$   | $R^2$ | $\beta$        | $p$   | $R^2$ | $\beta$         | $p$    | $R^2$ |
| Mathematics education | Basics                            | 0.06       | 0.313 | 0.05  | 0.08           | 0.191 | 0.16  | 0.14            | 0.025  | 0.10  |
|                       | Adaptivity                        | -0.02      | 0.767 | 0.05  | 0.05           | 0.411 | 0.15  | 0.17            | 0.006  | 0.11  |
|                       | Digital media                     | -0.03      | 0.678 | 0.05  | 0.01           | 0.931 | 0.15  | 0.02            | 0.741  | 0.08  |
|                       | Methodology                       | 0.09       | 0.169 | 0.06  | 0.04           | 0.541 | 0.15  | 0.11            | 0.082  | 0.09  |
|                       | Curricular aspects/Assessment     | 0.06       | 0.322 | 0.05  | 0.15           | 0.006 | 0.17  | 0.20*           | 0.001  | 0.12  |
|                       | Research on mathematics education | 0.08       | 0.138 | 0.06  | 0.17*          | 0.002 | 0.18  | 0.22**          | <0.001 | 0.13  |
| General pedagogy      | Adaptivity                        | 0.06       | 0.256 | 0.05  | 0.00           | 0.954 | 0.15  | 0.05            | 0.375  | 0.08  |
|                       | Structuring                       | 0.00       | 0.984 | 0.05  | 0.03           | 0.600 | 0.15  | 0.06            | 0.343  | 0.09  |
|                       | Classroom management/Motivation   | 0.02       | 0.673 | 0.05  | 0.04           | 0.527 | 0.15  | 0.03            | 0.589  | 0.08  |
|                       | Assessment                        | 0.10       | 0.076 | 0.06  | 0.09           | 0.110 | 0.16  | 0.01            | 0.880  | 0.08  |

All coefficients stem from separate models, including study semester, average grade in final secondary school examinations, teacher program, and one OTL subscale. The significance criterion was reduced to account for multiple significance testing

$\beta$  standardized regression coefficient

\* $p < 0.005$ , \*\* $p < 0.001$

a Bonferroni correction considering ten significance tests per noticing facet. While no significant effect was found for all OTL scales related to general pedagogy, both interpretation and decision-making were significantly predicted by the subscale “research on mathematics education,” and decision-making was further predicted by the subscale “curricular aspects/assessment.” The effect sizes for all significant coefficients were small. With respect to perception, no subscale showed any significant effect.



**Fig. 4** Factors within teacher education (individual prerequisites and use of OTL) predicting noticing skills (\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ); The correlation between noticing and the average grade is negative, since in Germany lower average grades indicate better performance

To report a final estimation of how much variance in the participants' noticing skills can be explained by individual prerequisites and use of OTL in the present study, a summarizing structural equation model was specified (see Fig. 4). This procedure is exploratory and can therefore not be transferred to other samples. For the model, perception, interpretation, and decision-making were used as indicators of noticing, and all variables that showed significant effects in the previous analyses were added as predictors, that is, the OTL subscales "research on mathematics education" and "curricular aspects/assessment," the average grade, the study semester, and the teacher education program. The OTL subscales were modeled as indicators of OTL in mathematics education as latent variable. The resulting model showed acceptable model fit ( $CFI=0.943$ ;  $RMSEA=0.063$ ;  $SRMR=0.055$ ;  $\chi^2=33.665$ ,  $df=16$ ,  $p<0.01$ ) and explained a considerable proportion of variance in pre-service teachers' noticing ( $R^2=0.226$ ).

## 6 Discussion

We investigated whether an established video-based test instrument, the TEDS-FU Video Test, originally developed to capture in-service mathematics teachers' noticing skills, could be used to measure pre-service teachers' noticing as a learning outcome of teacher education. The test was implemented in a survey of 313 pre-service teachers from different universities. We aimed to provide group-specific validity evidence by examining the internal test structure (RQ1) and the test's association with influential factors within teacher education including individual prerequisites and OTL (RQ2).

### 6.1 Measurement of perception, interpretation, and decision-making

With respect to our first research question, IRT scaling analysis revealed that the TEDS-FU Video Test provided a reliable measurement of the three noticing facets of perception, interpretation, and decision-making among the new target group. High correlations were found between perception and interpretation as well as for interpretation and decision-making, while perception and decision-making were only weakly correlated. This pattern corresponds to the theoretical assumptions on the structure of teacher noticing (Jacobs et al. 2010; Sherin et al. 2011b) and can thus be seen as validity evidence. Moreover, the correlations described are in line with the findings by Bastian et al. (2021), who investigated the TEDS-FU Video Test in a concurrent scaling analysis including pre-service teachers and in-service teachers.

Even when used for a pre-service teacher sample with limited formal access to practical teaching, the TEDS-FU Video Test provided a reliable measurement of the three noticing facets including decision-making skills. As novices, our target group can be assumed to have severe difficulties in quick decision-making (e.g., Carter et al. 1988; Stigler and Miller 2018) as they lack well-organized cognitive schemata and are not able to anticipate potential further courses of classroom events. However, they do not seem to be unfamiliar with classroom situations; their noticing abilities vary to a substantial degree, resulting in differentiated reliable measures.

We conclude from this that the transfer of this video-based noticing instrument as part of competence assessment into teacher education is possible.

## 6.2 Associations with factors within university teacher education

To interpret the test scores as a learning outcome of teacher education, evidence is required that the test scores are associated with relevant factors in teacher education including individual prerequisites and use of OTL. In our study, the average grade in the final secondary school examinations was the strongest predictor of noticing skills, having small to moderate effect sizes, which is in line with a previous meta-analysis highlighting (average) school grades as predictors of academic achievement (Trapmann et al. 2007). This association could be further explained by both the average grade in the final secondary school examinations and noticing being connected with information processing. This corresponds to the finding that the average grade showed a higher correlation with interpretation than with perception ( $z = 3.44$ ;  $p < 0.001$ ) and decision-making ( $z = 2.32$ ;  $p = 0.02$ ); this is possibly explained by interpretation being cognitively demanding and requiring knowledge when applying theories and concepts to observed instructional events. Overall, the effects of the average school leaving grade found in our study contradict the hypothesis by Stürmer et al. (2015), who assume that the average school grade in the final secondary school examinations is suitable to predict knowledge acquisition but not the application of knowledge into practice. The relationship between noticing and the average grade may, however, depend strongly on the used operationalization of noticing.

Although, on the theoretical level, noticing skills should be developed among pre-service teachers when engaging in teaching practice, no relationship between pre-service teachers' noticing and pedagogical experience was found in our study; this is in line with previous findings on the construct noticing (conceptualized as professional vision) measured by video-based tests (e.g., Stürmer et al. 2015; Todorova et al. 2017). Without explicit training, pre-service teachers may not automatically draw on their knowledge acquired during teacher education when they teach. Our findings suggest that targeted interventions are needed to support pre-service teachers in linking theory (e.g., principles of teaching and learning) and teaching practice (Stürmer et al. 2013; Weber et al. 2018).

For OTL in university teacher education, only few effects were found. OTL in mathematics education significantly predicted pre-service teachers' decision-making with a small effect size. This finding cannot be regarded as strong validity evidence. However, König et al. (2018) found no effect of overall OTL in mathematics education on pre-service teachers' pedagogical content knowledge. The authors assume that mathematics teacher education courses provide highly structured curricular requirements and therefore do not allow pre-service teachers a completely free choice of content during their study, thus reducing the variance of OTL measures and limiting effect sizes. Against this background and considering the small effect sizes in previous studies on the acquisition of teacher noticing (e.g., Wiens et al. 2013), it is encouraging that significant correlations between OTL and the TEDS-FU Video Test could be found at all.

König et al. (2018) found that only the subdimension “research on teaching mathematics” significantly predicted pedagogical content knowledge. In the present study, this subdimension predicted interpretation and decision-making, suggesting that this domain has diagnostic value. However, the significant effect of specific OTL in mathematics education on teacher noticing, including “research on teaching mathematics” and “curricular aspects,” can also be explained by highly constructive alignment. Especially, the content related to research, among others, includes theories on the development of mathematical competence (e.g., Bruner’s modes of representation) and the role of applying mathematics to real-word problems, both covered by the TEDS-FU Video Test. Moreover, the correlation between test scores and the subscale focusing on research might reflect that the development of this instrument, which was mainly conducted by researchers from mathematics education, was highly influenced by prominent mathematics educational theories and findings within this research area and reflect a common core of courses in mathematics education.

In contrast, the absence of general pedagogy OTL effects in our analysis may be interpreted regarding the strong focus of the test on teaching mathematics, even though general pedagogical issues are also included. However, with respect to the discourse on the effectivity of teacher education, it is also possible that for OTL in general pedagogy, the theoretical contents are not sufficiently related to practice, for example, using video clips or other forms of practical examples.

To sum up, the TEDS-FU Video Test allows for a reliable measurement of pre-service mathematics teachers’ noticing skills and is significantly associated with factors within teacher education, that is, individual prerequisites and OTL. Factors explained nearly 23% of the variance in noticing, which is comparable to previous studies (Stürmer et al. 2015; Todorova et al. 2017). However, the variance explained is particularly due to the average grade as an indicator of academic ability, while only few effects of OTL were shown, suggesting that test scores should not be interpreted as learning outcomes on the individual level (e.g., for individual diagnostics), but more measuring the effects of programs. This conclusion is also supported by the multilevel regression models reported, as OTL in mathematics education explained a considerable proportion of variance in interpretation test scores on program level.

### 6.3 Limitations and directions for future research

The following limitations should be considered. First, the analysis is based on a convenience sample, so the variance may be restricted due to selection effects. Also, it should be noted that the analyses are based on cross-sectional data. So, the effects of OTL and study semester cannot be interpreted as effects of development.

Exploratory analyses of the internal structure revealed that the two-dimensional model merging interpretation and decision-making shows better model fit than the three-dimensional model. This result is partly due to the comparably low reliability of decision-making suggesting that further efforts in test development would be helpful. It should be emphasized that the effect sizes for interpretation and decision making were slightly different, suggesting that the facets are separable. Moreover,

the multilevel analysis conducted suggested that the two facets differ regarding their relationships to OTL on the individual and the program level.

Although a substantial amount of variance in noticing was explained by the factors considered, there is still a high proportion of unexplained variance—especially regarding the perception facet. Future studies should therefore identify further influencing factors, such as motivational aspects (e.g., interest), the actual extent and perceived quality of specific learning opportunities as well as teaching experience. In addition, the implementation of measures for cognitive abilities could be helpful to understand, whether the effect of the average grade on noticing skills leads back to cognitive abilities.

The absence of effects for many OTL scales in the present study might be due to operationalizing OTL as a list of topics within a domain. Future studies should develop and implement more specific questionnaires to explore the extent to which representations of practice (e.g., video clips) are used for supporting pre-service teachers' skills in analyzing classroom situations. Furthermore, previous studies using multilevel modeling to explore the acquisition of professional knowledge during teacher education, found that the influence of OTL was greater on program level than on the individual level (e.g., König et al. 2017). In the present study, the influence of OTL might be underestimated, especially for interpretation, since for this facet, OTL appears to be a relevant predictor on the program level. Future studies should aim at acquiring samples that are appropriate for multilevel modeling.

**Supplementary Information** The online version of this article (<https://doi.org/10.1007/s11618-023-01159-7>) contains supplementary material, which is available to authorized users.

**Funding** This work was supported by the German Ministry of Education and Research [Bundesministerium für Bildung und Forschung, BMBF, grant numbers: 01PK19006A, 01PK19006B].

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Conflict of interest** J. Weyers, J. König, B. Rott, G. Greefrath, K. Vorhölder and G. Kaiser declare that they have no competing interests.

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

## References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.
- Bastian, A., Kaiser, G., Meyer, D., Schwarz, B., & König, J. (2021). Teacher noticing and its growth toward expertise: An expert–novice comparison with pre-service and in-service secondary mathematics

- teachers. *Educational Studies in Mathematics*, 110(2), 205–232. <https://doi.org/10.1007/s10649-021-10128-y>.
- Baumert, J., & Kunter, M. (2013). Professionelle Kompetenz von Lehrkräften. In I. Gogolin, H. Kuper, H.-H. Krüger & J. Baumert (Eds.), *Stichwort: Zeitschrift für Erziehungswissenschaft*. Wiesbaden: Springer. [https://doi.org/10.1007/978-3-658-00908-3\\_13](https://doi.org/10.1007/978-3-658-00908-3_13).
- Berliner, D.C. (1988). *The development of expertise in pedagogy*. American Association of Colleges for Teachers.
- Blömeke, S., Hsieh, F.-J., Kaiser, G., & Schmidt, W.H. (Eds.). (2014). *International perspectives on teacher knowledge, beliefs and opportunities to learn: TEDS-M results*. Wiesbaden: Springer.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R.J. (2015a). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13. <https://doi.org/10.1027/2151-2604/a000194>.
- Blömeke, S., Hoth, J., Döhrmann, M., Busse, A., Kaiser, G., & König, J. (2015b). Teacher change during induction: Development of beginning primary teachers' knowledge, beliefs and performance. *International Journal of Science and Mathematics Education*, 13(2), 287–308. <https://doi.org/10.1007/s10763-015-9619-4>.
- Bond, T., & Fox, C.M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences*. London: Routledge.
- Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education*, 39(3), 25–31. <https://doi.org/10.1177/002248718803900306>.
- Choy, B.H., & Dindyal, J. (2020). Teacher noticing, mathematics. In M. A. Peters (Ed.), *Encyclopedia of teacher education*. Singapore: Springer. [https://doi.org/10.1007/978-981-13-1179-6\\_241-1](https://doi.org/10.1007/978-981-13-1179-6_241-1).
- Christiansen, I.M., & Erixon, E.-L. (2021). Opportunities to learn mathematics pedagogy and learning to teach mathematics in Swedish mathematics teacher education: A survey of student experiences. *European Journal of Teacher Education*. <https://doi.org/10.1080/02619768.2021.2019216>.
- Dindyal, J., Schack, E.O., Choy, B.H., & Sherin, M.G. (2021). Exploring the terrains of mathematics teacher noticing. *ZDM—Mathematics Education*, 53(1), 1–16. <https://doi.org/10.1007/s11858-021-01249-y>.
- Doll, J., Buchholtz, N., Kaiser, G., König, J., & Bremerich-Vos, A. (2018). Nutzungsverläufe für fachdidaktische Studieninhalte der Fächer Deutsch, Englisch und Mathematik im Lehramtsstudium. Die Bedeutung der Lehrämter und der Zusammenhang mit Lehrinnovationen. *Zeitschrift für Pädagogik*, 64, 511–532. <https://doi.org/10.25656/01:22164>.
- van Es, E.A., & Sherin, M.G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, 10(4), 571–596.
- Floden, R. (2002). The measurement of opportunity to learn. In National Research Council (Ed.), *Methodological advances in cross-national surveys of education achievement* (pp. 231–266). National Academies.
- Gold, B., & Holodynski, M. (2017). Using digital video to measure the professional vision of elementary classroom management: test validation and methodological challenges. *Computers & Education*, 107, 13–30. <https://doi.org/10.1016/j.compedu.2016.12.012>.
- Goodwin, C. (1994). Professional vision. *American Anthropologist*, 96(3), 606–633.
- Gräsel, C. (2010). Stichwort: Transfer und Transferforschung im Bildungsbereich. *Zeitschrift für Erziehungswissenschaft*, 13(1), 7–20. <https://doi.org/10.1007/s11618-010-0109-8>.
- Hoth, J., Schwarz, B., Kaiser, G., Busse, A., König, J., & Blömeke, S. (2016). Uncovering predictors of disagreement: ensuring the quality of expert ratings. *ZDM—Mathematics Education*, 48(1–2), 83–95. <https://doi.org/10.1007/s11858-016-0758-z>.
- Jacobs, V.R., Lamb, L.L.C., & Philipp, R.A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, 41(2), 169–202.
- Jamil, F.M., Sabol, T.J., Hamre, B.K., & Pianta, R.C. (2015). Assessing teachers' skills in detecting and identifying effective interactions in the classroom. *The Elementary School Journal*, 115(3), 407–432. <https://doi.org/10.1086/680353>.
- Kaiser, G., & König, J. (2020). Analyses and validation of central assessment instruments of the research program TEDS-M. In O. Zlatkin-Troitschanskaia, H.A. Pant, M. Toepper & C. Lautenbach (Eds.), *Student learning in German higher education* (pp. 29–51). Wiesbaden: Springer. [https://doi.org/10.1007/978-3-658-27886-1\\_3](https://doi.org/10.1007/978-3-658-27886-1_3).
- Kaiser, G., Busse, A., Hoth, J., König, J., & Blömeke, S. (2015). About the complexities of video-based assessments: Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. *International Journal of Science and Mathematics Education*, 13(2), 369–387. <https://doi.org/10.1007/s10763-015-9616-7>.



- KMK (2019). Standards für die Lehrerbildung: Bildungswissenschaften. [https://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2004/2004\\_12\\_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf](https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf). Accessed 20. February 2022
- König, J., Tachtsoglou, S., & Seifert, A. (2012). Individuelle Voraussetzungen, Lerngelegenheiten und der Erwerb von pädagogischem Professionswissen. In J. König & A. Seifert (Eds.), *Lehramtstudierende erwerben pädagogisches Professionswissen: Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (pp. 234–283). Münster: Waxmann.
- König, J., Rothland, M., Darge, K., Lünemann, M., & Tachtsoglou, S. (2013). Erfassung und Struktur berufswahlrelevanter Faktoren für die Lehrerausbildung und den Lehrerberuf in Deutschland, Österreich und der Schweiz. *Zeitschrift für Erziehungswissenschaft*, 16(3), 553–577. <https://doi.org/10.1007/s11618-013-0373-5>.
- König, J., Ligtvoet, R., Klemenz, S., & Rothland, M. (2017). Effects of opportunities to learn in teacher preparation on future teachers' general pedagogical knowledge: Analyzing program characteristics and outcomes. *Studies in Educational Evaluation*, 53, 122–133. <https://doi.org/10.1016/j.stueduc.2017.03.001>.
- König, J., Doll, J., Buchholtz, N., Förster, S., Kaspar, K., Rühl, A.-M., Strauß, S., Bremerich-Vos, A., Fladung, I., & Kaiser, G. (2018). Pädagogisches Wissen versus fachdidaktisches Wissen? *Zeitschrift für Erziehungswissenschaft*, 21(3), 1–38. <https://doi.org/10.1007/s11618-017-0765-z>.
- Korthagen, F. A. (2010). How teacher education can make a difference. *Journal of Education for Teaching*, 36(4), 407–423. <https://doi.org/10.1080/02607476.2010.513854>.
- Kunina-Habenicht, O., Schulze-Stocker, F., Kunter, M., Baumert, J., Leutner, D., Förster, D., Lohse-Bossenz, H., & Terhart, E. (2013). Die Bedeutung der Lerngelegenheiten im Lehramtsstudium und deren individuelle Nutzung für den Aufbau des bildungswissenschaftlichen Wissens. *Zeitschrift für Pädagogik*, 59(1), 1–23.
- Lindl, A., & Krauss, S. (2017). Transdisziplinäre Perspektiven auf domänenspezifische Lehrerkompetenzen. In S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff & R.-H. Mulder (Eds.), *FALKO. Fachspezifische Lehrerkompetenzen* (pp. 381–438). Münster: Waxmann.
- Mason, J. (2002). *Researching your own practice: the discipline of noticing*. London: Routledge.
- Mayr, J. (2010). Selektieren und/oder qualifizieren? Empirische Befunde zu guten Lehrpersonen. In J. Abel & G. Faust-Siehl (Eds.), *Wirkt Lehrerbildung? Antworten aus der empirischen Forschung* (pp. 73–89). Münster: Waxmann.
- Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1), 172–175. <https://doi.org/10.1037/0033-2909.111.1.172>.
- Mertens, S., & Gräsel, C. (2018). Entwicklungsbereiche bildungswissenschaftlicher Kompetenzen von Lehramtsstudierenden im Praxissemester. *Zeitschrift für Erziehungswissenschaft*, 21(6), 1109–1133. <https://doi.org/10.1007/s11618-018-0825-z>.
- Meschede, N., Steffensky, M., Wolters, M., & Möller, K. (2015). Professionelle Wahrnehmung der Lernunterstützung im naturwissenschaftlichen Grundschulunterricht: Theoretische Beschreibung und empirische Erfassung. *Unterrichtswissenschaft*, 43(4), 317–335.
- Muthén, B. O., & Muthén, L. K. (2006). *MPlus (Version 4.2) [Computer software]*. 1998–2006
- Neuweg, G. H. (2015). Kontextualisierte Kompetenzmessung: Eine Bilanz zu aktuellen Konzeptionen und forschungsmethodischen Zugängen. *Zeitschrift für Pädagogik*, 61(3), 377–383. <https://doi.org/10.25656/01:15368>.
- Santagata, R., König, J., Scheiner, T., Nguyen, H., Adleff, A.-K., Yang, X., & Kaiser, G. (2021). Mathematics teacher learning to notice: A systematic review of studies of video-based programs. *ZDM—Mathematics Education*, 53(1), 119–134. <https://doi.org/10.1007/s11858-020-01216-z>.
- Schmidt, W. H., Cogan, L., & Houang, R. (2011). The role of opportunity to learn in teacher preparation: An international context. *Journal of Teacher Education*, 62(2), 138–153. <https://doi.org/10.1177/0022487110391987>.
- Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in pre-service teachers. *American Educational Research Journal*, 51(4), 739–771. <https://doi.org/10.3102/0002831214531321>.
- Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (Eds.). (2011a). *Mathematics teacher noticing: Seeing through teachers' eyes*. London: Routledge.
- Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (Eds.). (2011b). Situating the study of teacher noticing. In M. G. Sherin, V. R. Jacobs & R. A. Philipp (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 3–13). London: Routledge.

- Star, J. R., & Strickland, S. K. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education*, 11(2), 107–125. <https://doi.org/10.1007/s10857-007-9063-7>.
- Stigler, J. W., & Miller, K. F. (2018). Expertise and expert performance in teaching. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt & A. M. Williams (Eds.), *The cambridge handbook of expertise and expert performance* (pp. 431–452). Cambridge: University Press. <https://doi.org/10.1017/9781316480748.024>.
- Stürmer, K., Seidel, T., & Schäfer, S. (2013). Changes in professional vision in the context of practice. *Gruppendynamik und Organisationsberatung*, 44(3), 339–355. <https://doi.org/10.1007/s11612-013-0216-0>.
- Stürmer, K., Könings, K. D., & Seidel, T. (2015). Factors within university-based teacher education relating to preservice teachers' professional vision. *Vocations and Learning*, 8(1), 35–54. <https://doi.org/10.1007/s12186-014-9122-z>.
- Tatto, M., Schulle, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher Education and Development Study in Mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics. Conceptual framework. Teacher education and development international study*
- Todorova, M., Sunder, C., Steffensky, M., & Möller, K. (2017). Pre-service teachers' professional vision of instructional support in primary science classes: How content-specific is this skill and which learning opportunities in initial teacher education are relevant for its acquisition? *Teaching and Teacher Education*, 68, 275–288. <https://doi.org/10.1016/j.tate.2017.08.016>.
- Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, 21(1), 11–27. <https://doi.org/10.1024/1010-0652.21.1.11>.
- Ulrich, I., & Gröschner, A. (Eds.). (2020). *Praxissemester im Lehramtsstudium in Deutschland: Wirkungen auf Studierende*. Wiesbaden: Springer.
- Voss, T., Kunina-Habenicht, O., Hoehne, V., & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft*, 18(2), 187–223. <https://doi.org/10.1007/s11618-015-0626-6>.
- Weber, K. E., Gold, B., Prilop, C. N., & Kleinknecht, M. (2018). Promoting pre-service teachers' professional vision of classroom management during practical school training: Effects of a structured online- and video-based self-reflection and feedback intervention. *Teaching and Teacher Education*, 76, 39–49. <https://doi.org/10.1016/j.tate.2018.08.008>.
- Wiens, P. D., & Gromlich, M. D. (2018). Five years of video-based assessment data: Lessons from a teacher education program. *Research & Practice in Assessment*, 13, 51–61.
- Wiens, P. D., Hessberg, K., LoCasale-Crouch, J., & DeCoster, J. (2013). Using a standardized video-based assessment in a university teacher education program to examine preservice teachers knowledge related to effective teaching. *Teaching and Teacher Education*, 33, 24–33.
- Wiens, P. D., Beck, J. S., & Lunsman, C. J. (2020). Assessing teacher pedagogical knowledge: the Video Assessment of Teacher Knowledge (VATK). *Educational Studies*. <https://doi.org/10.1080/03055698.2020.1750350>.
- Wolff, C. E., Jarodzka, H., & Boshuizen, H. P. A. (2021). Classroom management scripts: A theoretical model contrasting expert and novice teachers' knowledge and awareness of classroom events. *Educational Psychology Review*, 33(1), 131–148. <https://doi.org/10.1007/s10648-020-09542-0>.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Multi-aspect test software [Computer software]*. Australian Council for Educational Research.