Teaching and Teacher Education 122 (2023) 103970

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Review article

Measuring teacher noticing: A scoping review of standardized instruments

Jonas Weyers [a], Johannes König [a], Rossella Santagata [b], Thorsten Scheiner [c], Gabriele Kaiser [d, e, *]

[a] University of Cologne, Gronewaldstraße 2a, 50931 Köln, Germany
[b] University of California, Irvine, 3457 Education, 5500 Irvine, CA 92697, USA
[c] Australian Catholic University, 229 Elizabeth Street, Brisbane City, QLD 400, Australia
[d] Nord University, Universitetsalléen 11, 8026 Bodø, Norway
[e] University of Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany

HIGHLIGHTS

- Scoping review of 22 standardized test instruments that measure teacher noticing.
- Instruments are predominantly video-based and include operationalizations of different mental processes.
- Few instruments assess subject-specific noticing outside of mathematics teaching.
- Test quality varies considerably with no indication of internal consistency for some instruments.
- Validation by means of teacher knowledge, observed instructional quality and expert-novice comparisons is rarely conducted.

ARTICLE INFO

Article history:
Received 28 December 2021
Received in revised form
19 October 2022
Accepted 29 November 2022
Available online 9 December 2022

Keywords:
Teacher noticing
Teacher professional vision
Test instrument
Teacher expertise
Scoping review

ABSTRACT

This scoping review provides an overview of standardized instruments used to measure teacher noticing. A systematic literature search identified 37 publications in English-language peer-reviewed journals describing 22 different test instruments. Regarding the underlying conceptualization of noticing, instruments commonly distinguish mental processes (e.g., attending and interpreting) using heterogeneous nomenclatures and focus on various aspects of teaching. Regarding the test design, the instruments are predominantly video-based and vary considerably with respect to measurement approach and test requirements. High test quality was demonstrated for established test instruments. However, on a general level, desiderata became apparent regarding construct and criterion-related validity.

Contents

* Corresponding author.
E-mail address: gabriele.kaiser@uni-hamburg.de (G. Kaiser).

https://doi.org/10.1016/j.tate.2022.103970

1. Introduction[1]

During instruction, teachers are simultaneously confronted with large amounts of information from which they must identify relevant instructional events, reflect on them, and determine appropriate responses. This process is often referred to as *teacher noticing*, which is broadly defined as "specialized ways in which teachers observe and make sense of classroom events and instructional details" (Choy & Dindyal, 2020).

Teacher noticing is considered a central component of teachers' professional competence (Kaiser & König, 2019; Scheiner, 2016; Sherin, Jacobs, & Philipp, 2011a; Stahnke, Schueler, & Roesken-Winter, 2016) and considerable efforts have been made in recent years to develop various test instruments to measure noticing (e.g., *Kaiser, Busse, Hoth, König, & Blömeke, 2015; *Seidel & Stürmer, 2014). With increasing recognition of the importance and complexity of teacher noticing further instruments are still being developed.

The development of instruments to assess teacher noticing has posed significant challenges owing to noticing's volatile nature as an "in-the-moment-practice" (Jacobs, 2017, p. 273). Common approaches use classroom artifacts: for example, a teacher might watch a video clip of children discussing a mathematical problem and may then be asked to identify relevant utterances, interpret the children's mathematical understanding, and infer what the most appropriate response to the children would be (*Jacobs, Lamb, & Philipp, 2010). However, conceptualizations and operationalizations of noticing are heterogeneous across the various existing instruments.

As has been pointed out in the systematic literature review by König et al. (2022) most empirical studies on teacher noticing deploy qualitative approaches that provide detailed accounts of the nature and development of teachers' noticing. Standardized measurement of noticing, by contrast, is an important addition, as it enables the study of noticing in large samples of teachers and provides the basis of hypothesis testing. In this review, we therefore focus on standardized measurement approaches to noticing. These approaches allow researchers to empirically test theoretical assumptions, such as the conceptualization of noticing as a learning outcome of teacher education or as a correlate of professional knowledge. As the quality of research is limited by the quality of measures implemented (e.g., DeVellis, 2017), high-quality testing of noticing is needed to draw valid conclusions about the underlying theory.

Despite the recent publication of several systematic literature reviews on teacher noticing (Amador, Bragelman, & Castro Superfine, 2021; König et al., 2022; Santagata et al., 2021; Stahnke et al., 2016), no literature review to date has focused on standardized testing. An overview of existing noticing instruments will be useful, providing researchers with the necessary information to select instruments that are appropriate for their research goals, develop adequate measurement approaches and validation strategies, and identify areas for further test development. We conducted a scoping review, which is a suitable approach for this purpose (Munn et al., 2018; Noordink, Verharen, Schalk, van Eck, & van Regenmortel, 2021), to map existing instruments according to three main focal points. First, we describe the different conceptualizations of noticing that underlie the test instruments. Second, we focus on the test design, including the actual operationalization of teacher noticing. Third, we provide an overview of how researchers examined the quality of their instruments. Overall, this scoping review aims to identify research gaps and provide recommendations for future research.

2. Theoretical framework

2.1. Central conceptualizations of teacher noticing

The discourse on teacher noticing is characterized by various ways of describing and conceptualizing noticing, with teachers'

[1] Abbreviations: PVCM: Professional Vision of Classroom Management, PVIS: Professional Vision of Instructional Support, TEDS-FU: Teacher Education and Development Study—Follow Up, VAIL: Video Assessment of Interactions and Learning, SLR-PV: Professional Vision of Self Regulated learning, IRT: Item response theory, WLE: weighted likelihood estimation, EAP/PV: expected a posteriori estimation/plausible values.

perception as a core element (see Dindyal, Schack, Choy, & Sherin, 2021). In particular, terminological inconsistencies exist: for example, some researchers write about "teacher noticing," while others prefer the term "professional vision." Often, it remains unclear whether these two terms denote different constructs or represent similar concepts. In the section that follows, we aim to clarify this terminological issue by structuring the discourse according to four theoretical perspectives on teacher noticing, which we proposed in two systematic reviews: a socio-cultural perspective, a cognitive-psychological perspective, an expertise-related perspective, and a discipline-specific perspective (König et al., 2022; Santagata et al., 2021).

The concept of "professional vision" is central to the *socio-cultural perspective*, originating from the work of Goodwin (1994), who developed this concept with a focus on lawyers and archeologists. Goodwin (1994) described how professional vision—that is, a specialized way of seeing and understanding meaningful events in a professional context—is developed and shaped by social interaction in professional communities. Goodwin (1994) argued that professional vision as "the ability to see a meaningful event is not a transparent, psychological process, but instead a socially situated activity" (p. 606). Emphasizing the role of social interaction provides an important perspective on the acquisition of professional competence. However, owing to the focus on social interaction instead of the individual mind, this approach has been taken up only indirectly for the standardized testing of noticing.

Goodwin's (1994) general concept of professional vision was adapted for the teaching profession by Sherin and van Es (Sherin, 2001; Sherin, Russ, Sherin, & Colestock, 2008; Sherin & van Es, 2009), who focused on how participation in interactive video clubs shapes teachers' perception and sense-making of classroom interaction. This adaption of professional vision for the teaching profession, however, also prompted a shift in perspective: while ideas of socio-cultural embeddedness, which were immanent in the work of Goodwin (1994), were less emphasized, Sherin and van Es (with others) maintained a stronger focus on the mental processes in which teachers engage during instruction. The conceptualization of teacher noticing as a set of interrelated mental processes was characterized by König et al. (2022) as a *cognitive–psychological perspective* on noticing. The shift in perspective further entailed a shift in research methodology: while professional vision, as described by Goodwin (1994), lends itself to ethnographic or qualitative approaches, a focus on the individual teacher's cognitive processes may be regarded as a reference point for the standardized testing of noticing.

In their earlier work, Sherin and van Es (with others) referred to the "professional vision" construct, including the sub-processes of selective attention (also referred to as "noticing") and knowledge-based reasoning (also referred to as "interpreting") (Sherin, 2001; Sherin et al., 2008; Sherin & van Es, 2009). However, in their more recent work, they use "teacher noticing" to denote the overall construct, including subdimensions such as "attending to particular events" and "making sense of particular events" (Sherin, Jacobs, & Philipp, 2011, p. 5). Studies that were particularly important for measurement purposes foregrounded the term "professional vision" (e.g., *Seidel & Stürmer, 2014). Drawing on the work of Sherin and van Es, Seidel and Stürmer (2014) developed a standardized test instrument to assess teachers' professional vision, the Observer Research Tool, which has been influential for subsequent research. This has led to two consequences: first, professional vision and noticing are both commonly used to denote comparable sets of teachers' mental processes during instruction, thus suggesting interchangeability (see Huang, Miller, Cortina, & Richter, 2021). Second, the notion of "professional vision" has become increasingly independent of the socio-cultural perspective developed and elaborated by Goodwin (1994). Since the term "teacher noticing" is more commonly used in international research, in the following, we use "noticing" as a generic term and speak of "professional vision" only in reference to instruments for which this term is explicitly used by the authors who developed those instruments.

Aside from the two perspectives outlined so far, the roots of teacher noticing can also be found in teacher expertise research. This *expertise-related perspective* on noticing draws on the work of Berliner (1988), who studied the development of teaching skills from novice to expert. Although the term "noticing" is not used in this framework, studies on teacher expertise "can be regarded as precursors" (Lachner, Jarodzka, & Nückles, 2016, p. 198) because the concepts on which these studies have focused (e.g., interpreting and predicting classroom events) share similarities with the mental processes advocated by the cognitive–psychological perspective. For example, Sabers, Cushing, and Berliner (1991) demonstrated that expert teachers outperform novices with respect to their perception, monitoring, and interpretation of classroom events. Regarding research methodology, the expertise-related perspective emphasizes inter-individual differences in teachers' noticing skills, which is a precondition for the development of standardized measures.

Finally, research on teacher noticing was influenced by a fourth approach, characterized as a *discipline-specific perspective*. This approach was developed by Mason (2002), who understood noticing as a discipline in which teachers engage to enhance their sensitivity to classroom events. Teacher noticing is thus regarded as a "collection of practices designed to sensitize oneself so as to notice opportunities in the future in which to act freshly rather than automatically out of habit" (Mason, 2011, p. 61). As Mason (2021) notes, "the Discipline of Noticing […] is phenomenological in nature, being concerned with the lived experience of the practitioner" (p. 231) and, thus, does not directly relate to standardized measurement.

Although the four abovementioned theoretical perspectives share commonalities, their conceptualiziation of noticing differs with respect to focus and theoretical orientation, thereby implying different methodological orientations. In the context of standardized testing, which has been used in only a small proportion of studies on teacher noticing (König et al., 2022), the cognitive–psychological perspective is particularly relevant, since mental processes provide reference points for the operationalization of noticing. However, researchers have yet to reach a consensus on how many and which mental processes constitute noticing. On the one hand, van Es and Sherin (2002) distinguished "(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). On the other hand, *Jacobs et al. (2010) focused on teachers' professional noticing with respect to children's mathematical thinking and developed a model that included three sub-processes: *attending to* children's strategies, *interpreting* children's understanding based on the observed strategies, and *deciding how to respond*. *Kaiser et al. (2015) similarly conceptualized noticing as an interaction between perception, interpretation, and decision-making. This so-called PID-model is closely connected to Blömeke, Gustafsson, and Shavelson's (2015) conceptualization of competence as a continuum. For this model, the situation-specific skills—perception, interpretation, and decision-making—are conceptualized as mediator variables between cognitive and affective dispositions on the one hand and teaching performance on the other hand.

By contrast, *Seidel and Stürmer (2014) considered

"professional vision" to include two components: "noticing" and "reasoning." Noticing, as *Seidel and Stürmer (2014) understood it, denotes teachers' attention to relevant instructional events, taking into account goal clarity, teacher support, and learning climate. It is thus similar to the attending component of noticing in the conceptualizations described above. Reasoning, the second component of professional vision as conceptualized by *Seidel and Stürmer (2014), denotes teachers' interpretation of instructional events based on their professional knowledge. It is divided into three distinct but interrelated processes: (1) *describing* relevant instructional events based on professional knowledge; (2) *explaining* instructional events, including the connections between different events of the teaching-learning process; and (3) *predicting* the impact of instructional events on teaching and learning processes (Seidel, Blomberg, & Stürmer, 2010; *Seidel & Stürmer, 2014).

To conclude, the terms "noticing" and "professional vision" must be considered when examining instruments. Furthermore, it is important to consider which mental processes are differentiated and operationalized by different instrument developers.

2.2. Measuring teacher noticing

Teacher noticing primarily refers to what teachers "notice" during instruction ("in-the-moment-noticing," Sherin, Russ, & Colestock, 2011) and how they deal with what they have noticed, which is comparable to Schön's (1983) concept of "reflection in action." Since instructional practice is highly complex and testing requires standardization of the testing situation, researchers may struggle to create representations of practice—for example, using video clips, transcripts of instructional practice, or student written work samples (e.g., Dreher & Kuntze, 2015a; *Jacobs et al., 2010).

Existing instruments may be characterized along several dimensions, such as the underlying theoretical framework—which includes the conceptualization of noticing (i.e., which mental processes were characterized) and the domain-specific focus (i.e., what aspects of teaching and learning are "noticed")—and the test design, including the stimulus material used (e.g., video clips of teaching practice) and the test items (e.g., open-ended or closed-ended questions). Regarding the *underlying theoretical framework*, researchers are first challenged to choose or develop an accurate conceptualization of noticing that includes how many and which mental processes—called *noticing facets* (e.g., perception, interpretation, and decision-making)—are measured and how these facets can be operationalized (e.g., *Kaiser et al., 2015). Furthermore, noticing is commonly measured with a *domain-specific focus* representing the aspects of teaching and learning in which teachers engage while taking the test. The focus may relate to subject matter content (e.g., *Steffensky, Gold, Holodynski, & Möller, 2015) or to general pedagogical aspects (e.g., *Seidel & Stürmer, 2014) or both (e.g., *Blömeke et al., 2015).

The *test design* refers specifically to the construction of tasks or items used to measure a construct of interest, combining them for a test instrument as well as to scoring procedures and test administration (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Measures used for teacher noticing commonly employ *stimulus material* consisting of artifacts of instructional practice—mostly video clips—in combination with writing prompts or closed-ended questions (*Jacobs et al., 2010; *Kaiser et al., 2015; *Seidel & Stürmer, 2014; *Steffensky et al., 2015). Underlying these approaches is the implicit assumption that teachers who watch videos of instructional practice engage in cognitive processes comparable to those that they encounter during their own instruction. Moreover, the development of an accurate scoring system—that is, defining correct and incorrect answers—poses particular challenges when measuring noticing, since scholars must define what constitutes "correct" or "incorrect."

Some studies have also applied specific technologies to examine teacher noticing during instruction—for example, small wearable cameras combined with subsequent recall interviews (Sherin, Russ, & Colestock, 2011). Eye-tracking was used to investigate teachers' gaze behavior while watching videos of instructional practice (see Grub, Biermann, & Brünken, 2020). Kosko, Heisler, and Gandolfi (2022) studied pre-service teachers' head movements while the teachers viewed a 360-degree video using a virtual reality headset. However, such methods cannot easily be applied to large samples and do not fully allow for standardization of the testing situation. Therefore, these approaches are not included within the scope of our review.

To conclude, the development of standardized noticing measures, particularly those based on video material, poses significant challenges for researchers (Jacobs, 2017; *Kaiser et al., 2015; Nickerson, Lamb, & LaRochelle, 2017). Thus, the provision of an overview of the underlying conceptualizations and the test designs of existing instrument may facilitate future test development. During the test development process, both the underlying conceptualization and the test design are closely connected to the assessment of test quality.

2.3. Test quality

For this scoping review, we focused on the two classical aspects of test quality: reliability and validity, which were also emphasized in the Standards for Educational and Psychological Testing (AERA et al., 2014). Drawing on classical test theory, reliability denotes the precision of a measurement indicated by "the correlation between scores on two equivalent forms of the test" (AERA et al., 2014, p. 33). In broader terms, reliability denotes "the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported" (AERA et al., 2014, p. 33), thus referring to a range of possible coefficients (e.g., Cronbach's alpha, generalizability coefficients etc.).

Building on reliability, "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). To investigate validity, it is thus necessary to clarify a test's intended interpretation and to collect evidence in support of this interpretation. Possible sources of evidence include the analysis of test content (with respect to the construct addressed), the response processes (e.g., cognitive processes while taking the test), the test's internal structure (e.g., using factor analysis), and relationships to other variables (e.g., test-criterion relationships) (AERA et al., 2014).

While AERA et al. (2014) understand validity as a "unitary construct" (p. 14), earlier conceptualizations differentiated three validity types, which are still commonly used in research: (1) content validity, requiring the test items to be an adequate sample of all possible items that measure the construct; (2) criterion-related validity (also called predictive validity), focusing on the empirical relationship between a measure and a criterion measure; and (3) construct validity, which requires the test score's correlation with other variables to be consistent with theoretical assumptions regarding the relationship between the construct measured and other constructs or measures (Cronbach & Meehl, 1955; DeVellis, 2017).

Regarding the quality of tests used to measure noticing, it is

worth examining which concrete operations or strategies are used to assess reliability and validity. An overview of existing approaches can help provide a guideline for providing validity evidence concerning existing and newly developed instruments and reveal desiderata in test validation.

3. Research questions

The present paper aims to provide a scoping review of existing standardized test instruments used to assess teacher noticing by addressing three research questions:

1. How was the noticing construct conceptualized, including the overarching concept, the mental processes (noticing facets) distinguished, and the domain-specific focus?
2. How were test instruments for teacher noticing designed with respect to stimulus materials, test items, scaling, and scoring?
3. How was test quality examined, specifically in relation to the required quality standards of reliability and validity?

The answers to these questions will enhance our scientific knowledge of the state of the field and help identify key research gaps—that is, areas in which further research with existing instruments or even the development of new test instruments is required.

4. Method

To address the questions raised above, we conducted a scoping review. Addressing exploratory research questions, scoping reviews encompass the mapping of existing evidence in a topic area based on a systematic literature search thereby identifying research gaps and allowing first insight into the field (see Colquhoun et al., 2014). Literature selection, data collection, and reporting were in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR; Tricco et al., 2018).

4.1. Selection process

We conducted a systematic literature search to identify relevant papers. We first searched using the terms "teacher* AND notice*" as well as "teacher* AND professional vision."[2] As described above, the terms "noticing" and "professional vision" are regularly used interchangeably or as closely related terms. We included both since we did not wish to exclude relevant literature for terminological reasons.

The search was conducted across five online databases (ERIC, PsycINFO, ScienceDirect, Scopus, and Web of Science) and considered the titles, abstracts, and keywords of the publications. No restrictions were placed on publication year or publication type. This procedure resulted in 7205 publications in June 2019 following the removal of duplicates.

To screen publication titles and abstracts, we applied the following three inclusion criteria: (1) publication in a peer-reviewed journal; (2) publication in English; and (3) explicit focus on teacher noticing in the publication. Articles not published in peer-reviewed journals ($n = 2831$) were excluded to ensure that only high-quality publications were considered; publications in languages other than English ($n = 962$) were excluded to ensure a high level of accessibility; and publications that did not focus on teacher noticing ($n = 3186$) were excluded to ensure that only publications relevant to our purposes were selected. This screening yielded a total of 226 peer-reviewed English-language journal articles focused on teacher noticing. Full-text versions were then retrieved and reviewed by the authors, and the publications' relevance was assessed. Publications had to meet the two following inclusion criteria: (1) relevance to the discourse on teacher noticing, and (2) use of standardized testing to assess teacher noticing. Publications in which teacher noticing was not a construct or phenomenon of interest in the full-text version ($n = 44$) and publications that did not use standardized tests to measure teacher noticing ($n = 145$) were excluded. AERA et al. (2014) defined a test as "a device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process" (p. 2). Thus, a measure was considered to be standardized testing if the score indicating the participants' noticing capability was computed based on a standardized procedure. This yielded 37 publications that were ultimately included in this review. Fig. 1 summarizes the selection process.

4.2. Data charting and synthesis of results

The first coding phase included the entire sample of articles on teacher noticing ($n = 182$), which were screened for relevance in the course of evaluating the 226 full-text versions and excluding 44 articles in which teacher noticing was not a construct or phenomenon of interest (see Fig. 1). This sample was used for a broader literature review on conceptualizations of noticing and research methods used to study it (König et al., 2022). For this purpose, a coding scheme focusing on conceptualizations and research methods was developed by reviewing a subsample of 20 articles. A first version of the coding scheme was applied to an additional 20 articles and revised as necessary. All 182 articles were coded according to this final version, including double coding for 20%. Coding was conceptualized as dichotomous, meaning that the coder had to determine whether the article included a particular item of information. Interrater reliability can be described as good ($M_{Kappa} = .72$, $Min = 0.35$, $Max = 1.0$, $SD = 0.19$). The coding team discussed unclear coding decisions for all articles. With respect to the present review focusing on test instruments, this first coding scheme was used to retrieve information concerning the study design (e.g., cross-sectional, pre-post), the sample surveyed (e.g., in-service teachers, pre-service teachers) and to identify those articles that included standardized testing of teacher noticing.

A second phase of coding was conducted for the present review, and included articles that reported on the standardized testing of teacher noticing ($n = 37$). It should be noted that these 37 articles describing test instruments were only a small proportion of the whole literature selection ($n = 182$) indicating that the theoretical framework and methodology had to be investigated in more detail. Based on a review of these articles, a second coding scheme was developed to focus on.

- the *construct being measured*, including
  - o overarching concept (e.g., noticing, professional vision),
  - o noticing facets (e.g., perception, interpretation, and decision-making), and
  - o domain-specific focus (e.g., student thinking, subject matter content);
- the *test design*, including
  - o stimulus material (e.g., video clips),
  - o item format and number (e.g., number of rating items),

[2] The truncation symbol was added to ensure that the search results included all possible word-endings, particularly plural forms and gerunds. Searches using only the term "vision" yielded numerous references that were irrelevant to this review. We therefore used the complete term "professional vision."

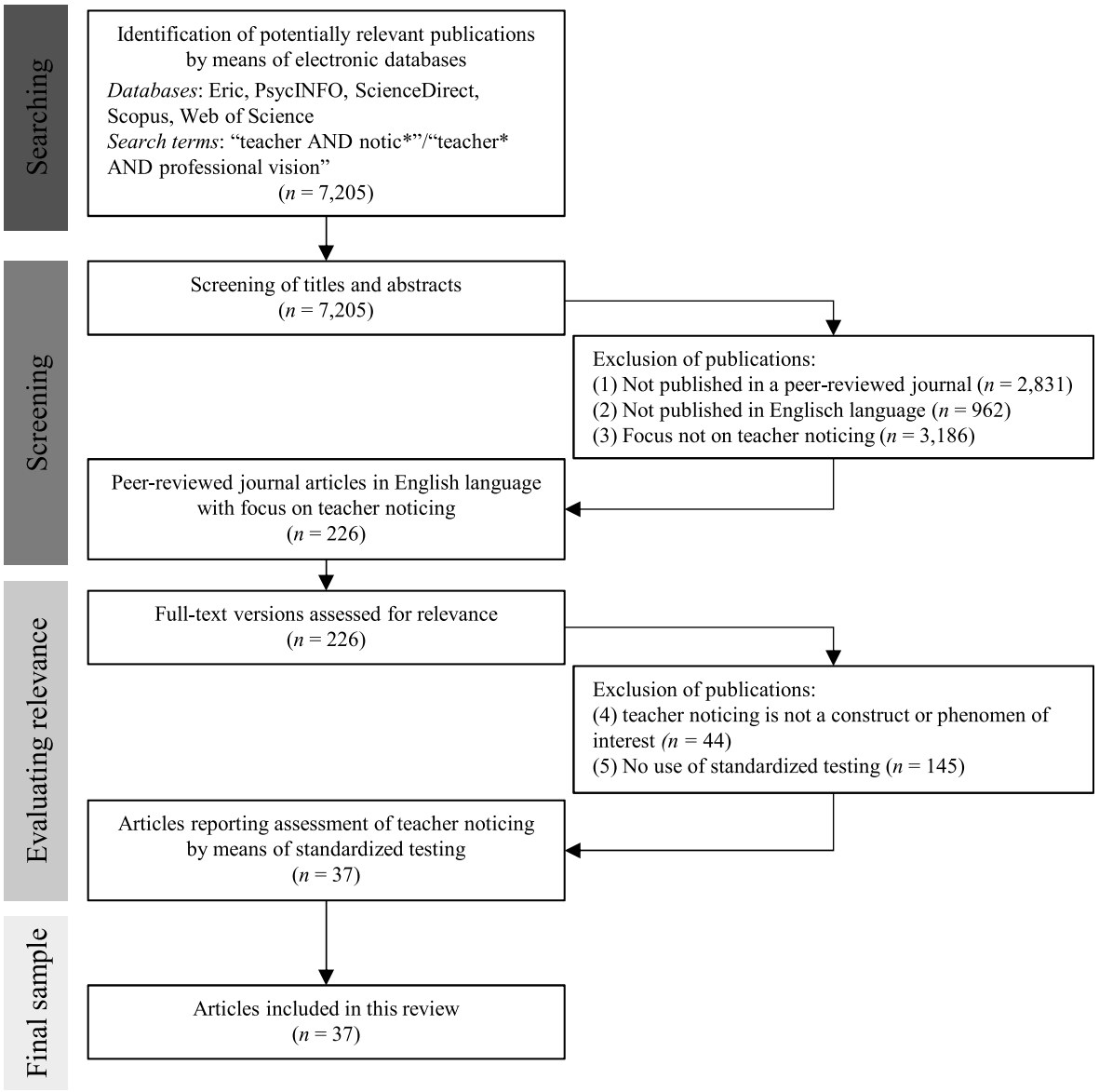

Fig. 1. Search and selection process.

- o scaling (e.g., mean scale, sum scale), and
- o scoring (e.g., scoring based on a coding manual);
- and the *test quality*, including
  - o reliability (e.g., internal consistency, interrater agreement) and
  - o validity (e.g, content validity, construct validity).

This coding scheme included dichotomous codes as well as open categories (e.g., nomenclature of noticing facets/domain-specific focus, values of coefficients etc.). Fifteen articles were double-coded, indicating good interrater reliability (M_{Kappa} = .84, *Min* = 0.46, *Max* = 1.0, *SD* = 0.13). The remaining articles were coded by one coder. For both coding phases—double and single coding—the research team discussed any discrepancies and unclear decisions. All coding categories from both coding manuals ultimately used for this review can be found in supplementary material A, including kappa values.

To provide an overview of the tests instruments identified in this review, charted data are presented in tables, including short presentations of each test instrument (see Table 1 and supplementary materials B and C). A synthesis of the results is provided in the text.

5. Results

5.1. Basic characteristics of the articles

Of the 37 articles included in this systematic literature review, the earliest was published in 2008. Most were published by European researchers (25), while the remainder were authored by researchers from North America (10) and Asia (2), including one collaboration between Chinese and European researchers (*Yang, Kaiser, König, & Blömeke, 2019). Cross-sectional designs (22), pre-post designs (15), and a longitudinal design with more than

Table 1
Overview of identified test instruments.

Test instrument	Description	Overarching concept[b]	Noticing facets Domain-specific focus
'Observer' Blomberg et al. (2011) *Seidel and Stürmer (2014) *Stürmer, Könings, & Seidel (2013) *Stürmer, Könings, & Seidel (2015) *Stürmer, Seidel, & Holzberger (2016) *Stürmer, Seidel, & Schäfer (2013)	The instrument was designed for pre-service teachers in different subjects. Using 112 rating items and six video clips (about 3.5 min)[a] of various subjects, participants are asked to agree/disagree with statements about observed instruction. Ratings are scored as correct if they match an expert rating.	Professional Vision	Knowledge-based reasoning • Description • Explanation • Prediction • Goal clarity • Teacher support • Positive learning climate
'Observer Extended' *Stürmer & Seidel (2015)	This is a modified version of the Observer, designed to survey pre-service teachers and teacher candidates during their induction phase. The instrument includes 10 video clips (about 3 min) and 41 rating items, and covers more areas of teaching than the original version.	Professional vision	Knowledge-based reasoning • Description • Explanation • Prediction • Goal setting • Orientation • Execution of learning activities • Evaluation of learning processes • Teacher guidance and support
TEDS-FU Video Test (primary) *Blömeke et al. (2015) *Kaiser et al. (2017)	The instrument was designed for early career primary teachers having originally participated in the study of TEDS-M for primary teachers. The instrument includes three scripted video vignettes (about 3.5 min) of primary school mathematics classroom instruction covering central topics in 3rd-grade mathematics. Participants work on 49 rating items and 38 open-response items. The scoring is based on an expert rating and by means of a detailed coding manual.	Noticing	• Perception • Interpretation • Decision-making • General pedagogy-related aspects • Mathematics instruction-related aspects
TEDS-FU Video Test (secondary) *Kaiser et al. (2015) *König et al. (2014) *Kaiser et al. (2017) *Yang et al. (2019)	The instrument is an equivalent version of the TEDS-FU Test (primary) focusing on secondary early-career teachers, who had participated in the study TEDS-M for secondary mathematics teachers. The video vignettes refer to central topics in school mathematics from years 8–10. The instrument includes 38 rating items and 36 open-response items.	Noticing	• Perception • Interpretation • Decision-making • General pedagogy-related aspects • Mathematics instruction-related aspects
'Video Case Diagnosis task' *Dalvi and Wendell (2017)	The instrument was used for pre-service primary teachers, engineering students, and teacher educators with experience in engineering and science education. It includes one video clip (6 min) of primary students working on an engineering design problem. Using four prompts, participants are asked to identify children's ideas and practices regarding science and engineering and to suggest responses to develop the children's understanding. The participants receive the video's transcript to work on these tasks. Based on an expert solution, coding rubrics were developed for scoring.	Noticing	• Noticing • Responding • Science ideas • Engineering practices
Multiple Representations Questionnaire Dreher and Kuntze (2015a) Dreher and Kuntze (2015b)	This instrument was designed for pre-service and in-service mathematics teachers. It includes four short transcripts of fictious teacher-student interaction, all focused on one specific aspect of school mathematics. For each vignette, the participants work on one writing prompt. Answers are scored as correct if the participants evaluate the teacher's response negatively and justify this evaluation by referring to a change in representations of fractional arithmetic.	Noticing	• Holistic approach (theme-specific noticing) • Multiple representations in mathematics classes
Noticing Measure by Jacobs et al. *Jacobs et al. (2010)	The instrument was designed for pre-service and in-service primary teachers as well as emerging teacher educators with different experience in children's mathematical thinking. Two assessments are combined: one is based on a video clip (9 min), the other contains three samples of students' written work. Both assessments refer to primary school mathematics classrooms and include three writing prompts: participants are asked to (1) describe, (2) interpret, and (3) respond to children's understandings. Each response is coded with respect to the degree (robust, limited, lack) of evidence for participants' noticing.	Noticing	• Attending • Interpreting • Deciding how to respond • Children's mathematical thinking (strategies and understanding)
Noticing Measure by Fisher et al. *Fisher et al. (2019)	The instrument was designed for pre-service teachers (primary education and special education), enrolled in an elementary mathematics methods course. It includes one short video (74 s) showing a group of 11 s-grade students who discuss the missing number in an equation. Participants are asked to describe what they observed, what the children understood, and what they would do next (three writing prompts). Answers are scored by means of decision trees.	Noticing	• Attending • Interpreting • Deciding how to respond • Children's mathematical thinking
Noticing Measure by Schack et al. *Schack et al. (2013) *Fisher et al. (2018)	The instrument was designed for pre-service elementary teachers enrolled in primary mathematics courses. It includes one short video clip (25 s) showing an interview with a single first-grade child. Referring to the child's thinking, participants are asked to describe, interpret, and decide how to respond (three writing prompts). The participants' answers are coded using a flowchart.	Noticing	• Attending • Interpreting • Deciding how to respond • Children's mathematical thinking (Stages of Early Arithmetic Learning)
Noticing Measure by Simpson and Haltiwanger *Simpson and Haltiwanger (2017)	The instrument was designed for pre-service mathematics teachers and includes three 12th grade student written work samples with each representing a different approach to a mathematics problem (algebra and function). Participants work on three writing prompts (describing,	Noticing	• Attend • Interpret • Respond • Student's mathematical thinking

(*continued on next page*)

Table 1 (*continued*)

Test instrument	Description	Overarching concept[b]	Noticing facets
			Domain-specific focus
	interpreting, deciding) for each work sample. The answers are coded based on the degree of evidence for noticing.		
Students' Course Outcomes *Johnson et al. (2019)	The participating pre-service teachers engage in online video-based learning environments that focus on professional noticing. Based on these learning environments, participants then work on performance-based tasks during their own instruction. Work on the performance-based tasks is evaluated by the researchers using Likert scales.	Noticing	• Holistic facet • Students' mathematical thinking
Monitoring Competence Assessment Tool *Kaendler, Wiedmann, Leuders, Rummel, and Spada (2016) *Wiedmann, Kaendler, Leuders, Spada, and Rummel (2019)	The instrument was designed for pre-service teachers and teacher candidates during their induction phase. It includes three short, scripted videos (about 1 min) showing groups of three students aged around 13 years. The videos depict students solving mathematics problems using collaborative, cognitive, and metacognitive activities. Participants rate 32 dichotomous items in terms of whether descriptive statements on students' activities are true. The answers are correct if they match an expert rating.	Professional vision	• Describing meaningful classroom events • Student interaction in collaborative learning settings
Comparative Judgment Instrument (primary) *Keppens et al. (2019)	The instrument was designed for pre-service primary teachers and contains 15 video clips (around 2 min) showing (inclusive) primary classrooms. Noticing is measured by comparative judgments: videos are presented pairwise and participants judge which of the two videos is better regarding two aspects of inclusive teaching (20 judgments). The participants score higher if their judgments deviate less from an expert rank order. Reasoning is measured using Likert scales; participants rate how important certain arguments were for their judgments (33 rating items).	Professional vision	• Noticing • Reasoning Inclusive classrooms • Teacher-student interaction • Differentiated instruction (Primary level)
Comparative Judgment Instrument (secondary) *Roose et al. (2018) *Roose, Vantieghem, Vanderlinde, and van Avermaet (2019)	The instrument was used for in-service teachers. It is equivalent to the Comparative Judgment Instrument (primary) but focuses (inclusive) on secondary classrooms. Reasoning is addressed by open-ended questions that, however, are not used for scoring.	Professional vision	• Noticing [Reasoning is not measured.] Inclusive classrooms • Teacher-student interaction • Differentiated instruction (Secondary level)
Noticing Measure of Adaptive Teaching *Kleinknecht and Gröschner (2016)	The instrument was designed for pre-service teachers (mathematics/science/informatics) and includes one video clip (3.5 min) showing mathematics instruction with a focus on adaptive strategies. The participants are asked to describe, and evaluate the adaptive instruction, and to create alternatives by means of three open-ended tasks. To measure 'selective attention', coders assess the number of perceived events; further aspects of knowledge-based reasoning are rated using a coding system.	Noticing	• Selective attention Knowledge-based reasoning • Reasoning process • Explanation/use of concepts • Dealing with negative events • Dealing with positive events Adaptive teaching
Assessment Scheme of Professional Vision of Self-Regulated Learning ('SRL-PV assessment scheme') *Michalsky (2014)	The instrument was designed for pre-service mathematics teachers and includes one video clip (25 min) of a high school mathematics lesson. The participants are prompted to specify the time stamp in the lesson when they notice that the teacher teaches self-regulated learning. They are further asked to describe and explain this situation and to predict how these instructional events will develop self-regulated learning. The participants' utterances are coded into four levels of professional vision depending on which processes are identifiable in the utterances.	Professional vision	• Noticing Knowledge-based reasoning • Describing • Explaining • Predicting Self-regulated learning • Direct delivery mode • Indirect delivery mode
Analyzing Teacher Moves Test *Scherrer and Stein (2013)	The instrument was designed for in-service mathematics teachers and includes one video (2.5 min) of secondary mathematics classroom discussion. Participants receive the transcript and ten (mostly open-ended) questions focusing what they paid attention to, what they appreciated and what alternative strategies they would propose. The answers are scored as correct if attention was on teacher-student interaction (unit of noticing), if specific codings were used (language use), and if opportunities to learn were related to teacher-student interaction. The number of possible points is not restricted.	Noticing	• Unit of noticing [similar to "attending"] • Language use [similar to "interpreting"] • Opportunities to learn [specific form of "interpreting"] • Teacher-student interaction in classroom discussions
Non-Interpretative Noticing Measure *Star and Strickland (2008)	The instrument was used for pre-service mathematics teachers and includes one video of a whole mathematics lesson (45 min). After watching the video, the participants work on 61 items of several formats. Items refer to clearly observable facts and thus do not require any interpretation (e.g., participants are asked to list as many names of students from the video as they remember).	Noticing	• Attending (no interpretation included) Instruction features • Classroom environment • Classroom management • Tasks • Mathematical content • Communication
Professional Vision of Classroom Management Test (PVCM Test) *Steffensky et al. (2015) *Gold and Holodynski (2017) *Weber, Gold, Prilop, and Kleinknecht (2018)	The instrument was designed for pre-service and in-service primary teachers. It includes four video clips (about 3 min) that depict extracts of primary science lessons. Using 47 rating items, the participants disagree/agree with statements referring to classroom management in the observed instruction. The ratings are scored as correct if they match an expert rating.	Professional vision	• Noticing • Interpretation Classroom management • Monitoring • Managing momentum • Rules and routines
Professional Vision of Instructional Support Test (PVIS Test) *Steffensky et al. (2015) *Meschede, Fiebranz, Möller, and	The instrument was used for pre-service and in-service primary teachers. It includes six video clips (about 3.5 min) which mainly show teacher-class interaction during primary science lesson. Similar to the PVCM Test, participants work on 68 rating items focusing instructional support. The ratings are scored as correct if they match an expert rating.	Professional vision	• Noticing • Interpretation Instructional support (in science classes) • Structuring

Table 1 (*continued*)

Test instrument	Description	Overarching concept[b]	Noticing facets
			Domain-specific focus
Steffensky (2017) *Todorova et al. (2017)			• Cognitive activation
Tagging Assessment *Theelen, Beemt, and Brok (2019)	The instrument was designed for pre-service teachers (various subjects) and includes three video clips (about 3.5 min) showing extracts of secondary instruction. The participants are asked to 'tag' the clips, i.e., note three to five important aspects about teacher-student relationship in the video. The 'tags' are then coded with respect to the analytical level: (1) descriptive, (2) evaluation, (3) analytic, and (4) prescriptive.	Professional vision	• Holistic approach • Interpersonal teacher behavior
'Video Assessment of Interactions and Learning' (VAIL) *Wiens and Gromlich (2018)	The instrument was originally developed for early childhood teacher but used for in-service and pre-service teachers from various domains. It includes three videos (about 2.5 min) of pre-school language arts classrooms. After each video, the participants are prompted to identify up to five teaching strategies and give specific examples from the video. The answers are scored by means of a coding manual. Each strategy-example pair is coded with respect to the identified strategies and examples, the match between strategy and example and the breadth of identified strategies. This results in 58 possible points for the version used *Wiens and Gromlich (2018).	Noticing	•Skill • Knowledge [Facets rather refer to the scoring method than to teachers' mental processes] Teaching strategies • Instructional supports • Classroom organization • Instructional supports

Notes. [a] The video durations given in parenthesis are the approximate duration per video included in the test. [b] It should be noted that the term "professional vision" does not indicate a socio-cultural perspective on noticing.

two measurements (*Stürmer, Seidel, & Holzberger, 2016) were reported.[3] The samples studied included pre-service teachers in 28 articles and in-service teachers in 15 articles, with seven articles including both pre-service and in-service teachers.

5.2. Identification of test instruments

The 37 papers included a total of 22 different test instruments, some of which were used in multiple papers. These test instruments are outlined in Table 1.[4]

Some of the identified test instruments relate to one another: this concerns the instrument "Observer" (Blomberg, Stürmer, & Seidel, 2011), for which an extended version ("Observer Extended") was reported by *Stürmer and Seidel (2015). Two instruments were derived from the project "Video-based lesson analysis: Early science," referred to as "Professional Vision of Classroom Management Test" (PVCM Test, *Gold & Holodynski, 2017) and "Professional Vision of Instructional Support Test" (PVIS test, *Todorova, Sunder, Steffensky, & Möller, 2017). Two further instruments come from the project "Teacher Education and Development Study—Follow Up" (TEDS-FU) and are called "TEDS-FU Video Tests" (primary and secondary) (*Blömeke et al., 2015; *Kaiser et al., 2015). Two instruments are from the project "Potential: Power to teach all" and labeled "Comparative Judgment Instruments" (*Keppens, Consuegra, Goossens, Maeyer, & Vanderlinde, 2019; *Roose, Goossens, Vanderlinde, Vantieghem, & van Avermaet, 2018). In addition, *Jacobs et al.'s (2010) data collection approach, which consists of three open-ended questions (describing, interpreting, and deciding how to respond), has been adopted by other researchers using different stimulus materials and coding procedures (*Fisher et al., 2018, 2019; *Schack et al., 2013; *Simpson & Haltiwanger, 2017). This led to similar but distinct instruments, which we labeled as "Noticing Measures" followed by the first author's name (see Table 1).

5.3. Conceptualizations of noticing

5.3.1. Noticing concept and noticing facets

Table 1 presents the overarching concepts and noticing facets distinguished for each test instrument. The overarching concept was "noticing" for 13 instruments and "professional vision" for nine instruments.

As the final column of Table 1 indicates, considerable heterogeneity emerged with respect to which noticing facets were addressed and how these facets were named. One approach to structuring the field is to assign the facets to one of three categories: (1) perceiving/attending, (2) reasoning/interpreting, and (3) deciding/responding (see Fig. 2). Regarding the instruments based on the noticing concept, for six instruments, noticing was found to include all three categories. Regarding the professional vision concept, the conceptualizations commonly focus on categories (1) and (2) (seven instruments).[5] For both versions of the Observer and for the "Assessment Scheme of Professional Vision of Self Regulated Learning" (SLR-PV) developed by *Michalsky (2014), the measurement of category (2) was further differentiated into description, explanation, and prediction.[6]

Some instruments are restricted to certain categories: for example, the "Non-Interpretative Noticing Measure" (*Star & Strickland, 2008) only addresses attending. For the Observer (*Seidel & Stürmer, 2014), only knowledge-based reasoning—that is, description, explanation, and prediction—is measured, although noticing was considered a subdimension of professional vision on the theoretical level. By contrast, holistic measurement approaches—wherein noticing is measured as a single construct with no distinction of processes—were used for only three instruments: the "Multiple Representations Questionnaire" (Dreher & Kuntze, 2015a), the "Students' Course Outcomes" by *Johnson et al. (2019), and *Theelen et al.'s (2019) "Tagging Assessment." Table 1 includes some less common conceptualizations, such as *Kleinknecht and Gröschner's (2016) "Noticing Measure of

[3] *Seidel and Stürmer's (2014) publication was double counted because it reported two studies.

[4] When a unique name for the instrument was used in the publications, this name was adopted for the present review (indicated by single quotation marks in Table 1). For the remaining instruments, we chose names that we felt represented the most salient features of the respective instruments.

[5] The nomenclature of the PVCM and PVIS Tests varies between publications. According to *Steffensky et al. (2015), professional vision includes "noticing" and "interpretation." For PVCM, *Gold and Holodynski (2017) included "description" and "interpretation."

[6] Only one score is computed for *Michalsky's (2014) SLR-PV assessment scheme, since the noticing facets are used in the sense of levels with description as the lowest and prediction as the highest level.

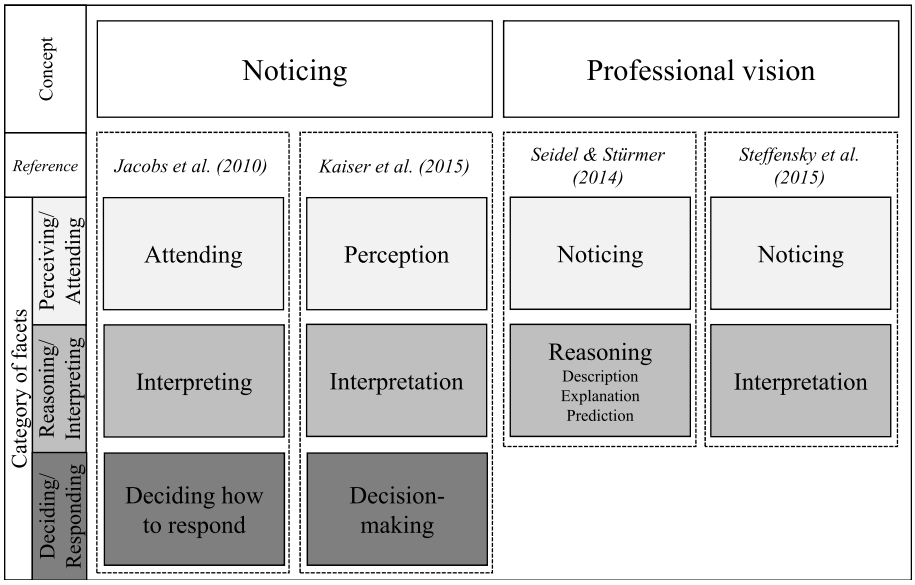

Fig. 2. Selected conceptualizations underlying test instruments.

Adaptive Teaching," *Scherrer and Stein (2013) "Analyzing Teacher Moves Test" (*Scherrer & Stein, 2013), and the "Video Assessment of Interactions and Learning" (VAIL; *Wiens & Gromlich, 2018).

5.3.2. Domain-specific focuses

The test instruments typically focus on several domains (i.e., aspects of teaching and learning). The categories identified were student thinking (10 instruments), subject matter content (12), classroom management (6), and general pedagogy (13). The final column of Table 1 lists each instrument's focus. The 12 instruments relating to subject content focused on mathematics (9 instruments), science (2), and mathematics and science (1).

The domain-specific focus also depends on the subject and grade level to which the stimulus material used relates. Most test instruments (16) use stimulus material from a single subject: mathematics (12), science (3), and language (1),[7] while six instruments refer to two or more subjects. For two instruments, the subjects were not specified. In terms of school level, primary (8) and secondary (13) levels are frequently addressed, while the VAIL (*Wiens & Gromlich, 2018) is the only instrument that includes material from pre-school language art classrooms.

5.4. Test design

In the following subsections, we focus on general trends in test and item design. Further details on the stimulus material and the items used can be found in supplementary material B and Table 1.[8]

5.4.1. Stimulus material

The vast majority of 20 instruments use video material, while only the Multiple Representations Questionnaire (Dreher & Kuntze, 2015a) includes written vignettes, and only *Simpson and Haltiwanger (2017) Noticing Measure includes written samples of student work. The video material used generally consists of authentic classroom practice, whereas scripted video vignettes are used only for the TEDS-FU Video Tests (*Kaiser et al., 2015) and the "Monitoring Competence Assessment Tool" (*Kaendler et al., 2016). The number of video clips ($Min = 1$, $Max = 15$) and the length of video clips ($M = 6$ min 36 s, $SD = 11$ min 15 s) vary considerably between instruments. The use of three to six clips of 2 to 3 min duration is the most common approach.

5.4.2. Item format and item design

The choice of item format is a critical aspect of test development (see *Kaiser et al., 2015). We distinguished between open-response items, dichotomous items, rating items, and comparative judgments. Most instruments (17) contain only one item type: open-response items (11), rating items (4), dichotomous items (1), and comparative judgments (1). Both TEDS-FU Video Tests combine rating and open-response items, whereas *Keppens et al.'s (2019) Comparative Judgment Instrument (primary) includes comparative judgments and rating items. The "Non-Interpretative Noticing Measure" (*Star & Strickland, 2008) uses multiple item formats. Item counts vary from three open-response items to 112 rating items ($M = 39.44$, $SD = 33.63$).

To gain a deeper understanding of what teachers are asked to do during the assessment, we analyzed the sample items provided (see supplementary material C for details) in terms of item design principles. Rating items typically assess the extent to which the individual agrees with statements regarding observed instructional practice. This approach was used for the two Observer instruments (Blomberg et al., 2011; *Stürmer & Seidel, 2015), the two TEDS-FU Video Tests (*Blömeke et al., 2015; *Kaiser et al., 2015), and the PVCM and PVIS Tests (*Steffensky et al., 2015). Depending on the respective noticing facet, the statements are descriptive (e.g., "The teacher clarifies what the students are supposed to learn") or require an explanation (e.g., "The students have the opportunity to activate their prior knowledge of the topic") or prediction (e.g., "The students will be able to align their learning process to the learning objective") (Observer; Blomberg et al., 2011). The Monitoring Competence Assessment Tool (*Kaendler et al., 2016) adopts a similar approach, measuring the capacity to describe meaningful teaching events using dichotomous items (e.g., "Group members ask each other

[7] The frequencies of subjects included in the stimulus materials differ from the domain-specific focuses, since stimulus material from specific subjects can be used without addressing subject matter content.

[8] *Johnson et al. (2019) provided limited information about the measurement procedure. This instrument (Students' Course Outcomes) was therefore not included in the test design analyses.

questions when they do not understand something," true/false). For the TEDS-FU Video Test (secondary), *König et al. (2014) further distinguished rating items assessing precise perception (e.g., "The teacher presents the lesson's task visually AND acoustically").

For the two Comparative Judgment Instruments (*Keppens et al., 2019; *Roose et al., 2018), the comparative judgment method is used to measure noticing, which is understood as an attentional sub-process of professional vision. The version that applies to primary classrooms (*Keppens et al., 2019) includes rating items to capture the reasoning process. Test participants rated how important various arguments were to their previous judgments, with higher importance corresponding to higher reasoning scores.

Open-response items prompt test participants to describe, interpret, or generate responses to aspects of the stimulus material (e.g., "Please describe in detail what you think each child did in response to this problem," *Jacobs et al., 2010). In addition, items in this format may require participants to apply their knowledge of concepts and theories to the stimulus material. For example, the VAIL (*Wiens & Gromlich, 2018) asks participants to identify five instructional strategies from the video clip and provide a specific example for each strategy. The TEDS-FU Video Test (secondary) (*Kaiser et al., 2015) includes an item that asks test participants to describe the mathematical solution approaches of three pairs of students and explicitly targets the corresponding academic expressions (enactive-iconic-symbolic).

*Theelen et al.'s (2019) Tagging Assessment adopted a less typical approach: participants were asked to note three to five aspects from each video that they considered relevant to the teacher–student relationship.

5.4.3. Scoring and scaling

For open-ended item formats, test scores are assigned based on a coding scheme or manual (15 instruments). Expert responses are used to validate this scoring method (*Dalvi & Wendell, 2017; *Kaiser et al., 2015). For almost all instruments containing closed item formats, test scores were determined by comparing participants' ratings with those of a sample of experts. Table 1 provides brief descriptions of the scoring procedures for each test instrument.

Most instruments use scales based on classical test theory (e.g., sum or mean scales). Three instruments determine the test score with a single rating of one open-ended response. For six instruments, more sophisticated procedures based on item response theory (IRT) were used to estimate test scores.

5.5. Test quality

5.5.1. Reliability

Of the reliability measures based on classical test theory, internal consistency represents almost the only measure used in the present selection, with the exception of one study that reported retest reliability (Observer; Seidel & Stürmer, 2014). Cronbach's α is calculated by seven instruments and shows high reliability ($M = 0.85$, $Min = 0.64$, $Max = 0.98$). A summary of the reliability coefficients can be found in supplementary material D, Table 1.

Reliability measures based on IRT are reported for six instruments, including weighted likelihood estimation (WLE), expected a posteriori estimation/plausible values (EAP/PV), and scale separation reliability: both TEDS-FU Video Tests (*Blömeke et al., 2015; *Kaiser et al., 2015), both versions of the Observer (Blomberg et al., 2011; *Stürmer & Seidel, 2015), and both Comparative Judgment Instruments (*Keppens et al., 2019; *Roose et al., 2018). For three additional instruments, error variance due to nesting of items in the video clips was estimated using generalizability theory (Monitoring Competence Assessment Tool; *Wiedmann et al., 2019) and omega hierarchical (PVCM and PVIS Tests; *Steffensky et al., 2015). None of the above coefficients were reported for 11 instruments. A measure of interrater reliability was calculated for 9 of these instruments, while no reliability measure was found in the publications for the remaining two instruments.

5.5.2. Validity

Regarding the traditional classification of content, construct, and criterion-related validity, a validity type was coded as explicitly addressed if authors used the technical term to describe their procedure or if clearly attributable measures were reported.[9] While content validity (14 instruments) and construct validity (8) are frequently considered, investigations of criterion-related validity are rare (3).

Content validity was primarily assessed by asking experts about the validity of the items (12 instruments) and the stimulus material (11). For example, the experts assessed whether the video material was authentic and relevant to the domain-specific focus and depicted frequent and relevant classroom events (*Blömeke et al., 2015; *Schack et al., 2013; *Seidel & Stürmer, 2014). This approach goes hand in hand with the selection of appropriate video material (e.g., *Gold & Holodynski, 2017). In addition, experts rate the relevance of items and provide answers themselves, which are used to create a master rating with sufficient agreement among experts (e.g., *Kaiser et al., 2015).

Construct validity is commonly addressed by examining the internal structure of a test using factor analysis or IRT modeling (6 instruments). Group comparisons are reported as a measure of construct validity for Schack et al.'s Noticing Measure (see *Fisher et al., 2018) and the "Video Case Diagnosis task" (*Dalvi & Wendell, 2017).

Supplementary material E provides an overview of studies examining the internal structure and dimensionality of the test instruments. Three publications show that different domain-specific focuses of noticing may be distinguished (*Keppens et al., 2019; *Steffensky et al., 2015; *Todorova et al., 2017). Two publications favored a model with different factors for the different noticing facets (*König et al., 2014; *Seidel & Stürmer, 2014). In three publications, the authors advocated a unidimensional structure of noticing (*Gold & Holodynski, 2017; Jamil, Sabol, Hamre, & Pianta, 2015; *Meschede, Steffensky, Wolters, & Möller, 2015).

Relations to other variables with correlation or regression are reported in 11 papers (see supplementary material F). Although these papers do not explicitly target test validation, the results may be interpreted in the context of validity. Six publications relate noticing to professional (declarative) knowledge and demonstrate significant correlations between 0.25 and 0.56 (Dreher & Kuntze, 2015b; *Gold & Holodynski, 2017; *Kaiser et al., 2017; *König et al., 2014; *Meschede et al., 2017). Two publications show that noticing is related to teachers' beliefs (*Meschede et al., 2017; *Roose et al., 2019). *Blömeke et al. (2015) also examined the relationship between noticing, knowledge, and beliefs but compared teachers' profiles instead of using correlation. Using latent class analysis, they found that teachers with favorable knowledge and belief profiles attain higher noticing scores.

Six publications relate teacher noticing to aspects of (teacher) education and professional experience and report significant effects (*Keppens et al., 2019; *Stürmer et al., 2015). However, three

[9] For validity analysis, we checked whether publications on test development of the included instruments that were not part of the selection were cited. Three publications were included: Jamil et al. (2015) for the VAIL, *Meschede et al. (2015) for the PVIS Test, and Seidel et al. (2010) for the Observer.

publications found no significant effects for teaching experience or length of internship in school (*Roose et al., 2019; *Stürmer et al., 2015; *Todorova et al., 2017). Regarding high school grade point average, the Observer and the PVCM Test show no significant effects (*Stürmer et al., 2015; *Todorova et al., 2017), while a significant effect was observed for the VAIL (*Wiens & Gromlich, 2018).

Significant differences in test scores among groups with different levels of expertise demonstrate the test's sensitivity and indicate construct validity (Cronbach & Meehl, 1955). Thirteen publications report group comparisons; however, most do not focus on validity. Three publications draw comparisons between experts (e.g., teacher educators) and novices, and five publications draw comparisons between in-service and pre-service teachers. Six publications compare different groups of pre-service teachers (e.g., bachelor and master students), while four publications report comparisons among in-service teachers. The group comparisons are summarized in supplementary material G. Overall, the results demonstrate that the instruments have high sensitivity to different levels of expertise with varying effect sizes ($d_{Min} = .19$, $d_{Max} = .84$).

Some operations used to address construct validity can also be used to demonstrate criterion-related validity. For two instruments, comparisons between in-service and pre-service teachers are interpreted as evidence of criterion-related validity (PVCM and PVIS Tests; *Gold & Holodynski, 2017; *Meschede et al., 2015). For the VAIL only, criterion-related validity has been investigated by examining the correlation between test scores and observed instructional quality (Jamil et al., 2015).

6. Discussion

Using a scoping review approach, this study provides an overview of existing standardized instruments for studying teacher noticing, thereby identifying research gaps in this area. Based on a sample of 37 articles published between 2008 and 2019, we identified 22 different test instruments and examined (1) the theoretical conceptualization of noticing, (2) the test design, and (3) the test quality.

6.1. Summary of main results

Most instruments differentiate noticing into distinct mental processes while more holistic approaches are rare. The domain-specific focus of noticing varies considerably between instruments, with mathematical aspects predominantly investigated in subject-specific noticing.

In terms of test design, most instruments include video material from classroom practice to elicit noticing, typically using one to six video clips of up to 5 min in duration. The amount of video material, item format, item formulation, and number of items vary considerably, with the latter ranging from three writing prompts to more than 100 rating items. Test scores are commonly determined by comparing participant responses to an expert solution for closed-ended item formats or using a coding scheme for open-response items. Resulting test scores ranged from single values and mean scales up to IRT estimations for a small number of instruments.

Regarding test quality, high reliability scores were reported for around half of the instruments, with no reliability measures reported for only a few instruments. Validity examination was guided by the traditional division into content, construct, and criterion-related validity, with the latter rarely examined. Internal structure was analyzed by considering the different noticing facets as well as content-specific aspects, yielding heterogeneous results with regard to the structure of noticing, including its sub-processes. Few studies have investigated noticing in relation to knowledge and beliefs. Sensitivity to differences between groups with different levels of expertise is more frequently confirmed. However, experts and novices are rarely compared.

6.2. Conceptualizations

As the results suggest, the conceptual heterogeneity within the discourse on teacher noticing is equally evident in the field of standardized testing. With regard to the overarching concept, the instruments in the present selection were assigned to either “noticing” or “professional vision.” Although the instruments based on these two concepts differ in terms of domain-specific focuses and measurement approaches (supplementary material C, Tables 2, 3 and 4), the different terms used (i.e., noticing and professional vision) should not obscure the fact that the underlying constructs, including their sub-processes (“noticing facets”), are strikingly similar. As outlined in the theory section, it should be noted again that the term “professional vision” refers more strongly to a set of mental processes, however retaining the original idea of professional vision as specialized way of seeing and understanding events in a professional context. Differences between the constructs measured by instruments relate to the inclusion of a teacher's response, which can be found for noticing instruments. By contrast, the prediction facet is only captured by professional vision instruments. However, the decision-making facet of noticing in *Kaiser et al. (2015) includes anticipation, which is similar to the prediction facet of professional vision in *Seidel & Stürmer's (2014) study.

Another conceptual difference between noticing and professional vision in the context of standardized testing is that the perceiving/attending facet of noticing is further divided into (selective) attention and description in professional vision. However, this difference is not necessarily relevant at the empirical level. For example, the Observer (*Seidel & Stürmer, 2014) does not explicitly capture noticing as an attentional sub-process of professional vision but limits itself to capturing description. The attending facet of noticing can be measured by asking participants to describe what they observe (e.g., *Jacobs et al., 2010).

6.3. Inconsistencies in the measurement approaches

The following test instruments have been developed, characterized by an elaborate reliability investigation and a comprehensive validation procedure: the Observer (Blomberg et al., 2011), the PVIS and PVCM Tests (*Steffensky et al., 2015), the two Comparative Judgment Instruments (*Keppens et al., 2019; *Roose et al., 2018), the TEDS-FU Video Tests (primary and secondary) (*Blömeke et al., 2015; *Kaiser et al., 2015), the VAIL (*Wiens & Gromlich, 2018), and the Monitoring Competence Assessment Tool (*Kaendler et al., 2016).

For the Observer, the TEDS-FU Video Tests, and the Comparative Judgment Instruments, test quality is examined by means of IRT. The Observer is also characterized by a validation procedure including a survey on the appropriateness of the video clips, the examination of the factor structure, and the investigation of repeated measurement effects (*Seidel & Stürmer, 2014). For the TEDS-FU Video Tests, instructional practices were scripted and videotaped to ensure a high density of relevant instructional events on key topics in school mathematics (*Kaiser et al., 2015). The PVCM Test (*Gold & Holodynski, 2017) is distinctive in that it examines factor structure using a bifactor model that takes into account that test items are nested within video clips. The PVCM and PVIS Tests are both used to compare differences between in-service and pre-service teachers, including the investigation of measurement invariance (*Gold & Holodynski, 2017; *Meschede et al., 2017). The VAIL is distinguished by its criterion-related validation

with observed instructional quality (Jamil et al., 2015). *Wiedmann et al. (2019) implemented analyses based on generalizability theory, offering a promising approach to control for measurement error caused by video clips for the Monitoring Competence Assessment Tool. Finally, the Comparative Judgment Instruments provide an alternative that assesses noticing holistically (*Keppens et al., 2019; *Roose et al., 2018).

In contrast to the measurement procedures described above, no reliability measures (other than interrater reliability) were provided for a large proportion of the instruments. Similarly, for some instruments, validity evidence was only reported selectively, for example focusing exclusively on group comparisons. This finding indicates that the available methods for ensuring psychometric quality are not yet sufficiently used for some instruments.

Measurement approaches are inconsistent with respect to operationalizing noticing facets, particularly regarding the category of perceiving/attending. *Gold and Holodynski (2017) noted that measuring noticing (understood as selective attention) using standardized items is difficult because the item text directs attention. Also, for the Observer, noticing (understood as selective attention) is located in the process of video selection by the research team, whereas the rating items only assess reasoning (*Seidel & Stürmer, 2014). In the TEDS-FU Video Tests, perception is restricted to clear perceptual incidents and measured using both rating items and open-ended questions, while other processes, such as interpretation, are measured using open-response items (*Kaiser et al., 2015). Another approach to measuring noticing as an attentional sub-process of professional vision is the use of comparative judgments, even though a judgment about teaching quality certainly involves interpretation.

Finally, test instruments differ in the degree of declarative knowledge required to achieve a high test score. One group of instruments, often based on closed-ended questions, requires an estimation of the focused aspects of teaching quality, which is supposed to be knowledge-based. By contrast, other instruments, often based on open-response items, target the application of explicit, recallable knowledge, including technical terms. This suggests an overlap between noticing measures and contextualized testing of teachers' professional knowledge.

6.4. Implications and directions for future research

In addition to providing an overview of available instruments, this review aimed to identify research gaps that can be considered reference points for further research. First, the results suggest that instruments may be developed to assess further aspects of noticing. For example, the testing of subject-specific noticing is currently limited mainly to the subject of mathematics. Given that situated approaches to teacher competence, which have commonalities with the concept of teacher noticing, are gaining weight, researchers from other disciplines might consider whether the standardized testing of subject-specific noticing can enrich the competence assessment in their field. A first approach focuses on the subject-specific noticing of biology teachers (Kramer et al., 2020). Similarly, the decision-making facet was only rarely operationalized using high-quality testing. Since decision-making can be considered a central mediator to teachers' behavior during instruction, further test development should emphasize this facet.

Moreover, central empirical issues concerning construct and criterion-related validation have rarely been addressed, including the differences between experts and novices; the development of noticing during teacher education and the professional career; the relationship between noticing and teacher cognition, including knowledge and beliefs; and the study of criterion measures, such as observed teaching quality or student learning progress. Given that (construct) validity is closely related to the construct's theoretical conceptualization, to address these desiderata, the theoretical foundation of noticing must be strengthened. In this context, an important theoretical reference point is the expertise approach (Berliner, 1988; Sabers et al., 1991), which could stimulate future studies to examine the differences between experts and novices and the development from novice to expert. Blömeke et al. (2015) offer another compatible framework that considers teacher noticing described as situation-specific skills as a mediator between cognition and performance. This approach has not been adequately explored empirically and could stimulate the investigation of noticing as a correlate of teacher cognition and teaching performance. However, using these theoretical approaches for test development and validation should be critically evaluated by researchers since unilateral theoretical approaches might bias research results. For example, research based on a noticing measure that was constructed to be a correlate of professional knowledge might lead to overestimating the relevance of specific parts of teachers' professional knowledge such as declarative knowledge acquired from teacher education.

Another promising approach to construct validation, which received scant attention in our literature selection, is analysis of the correlation between different noticing measures. On the one hand, this may include the correlation between two or more standardized measures targeting at noticing (e.g., the Observer and the VAIL). On the other hand, it may be of interest to determine whether standardized noticing test scores are associated with the other modalities of measurement outlined above, including (mobile) eye-tracking, data retrieved from small wearable cameras, or the investigation of head movements while viewing 360-degree videos. In this context, it should be highlighted that no multitrait-multimethod analysis was reported in the selected literature, which would have allowed the effects of different measurement approaches (e.g., item formats) on the measurement of similar or distinct constructs to be taken into account.

The finding that the validation strategies mentioned above were rarely addressed by no means implies that each of these strategies should be used for each test instrument. By contrast, a stronger orientation toward the testing standards (AERA et al., 2014) might help advance the research field. This includes the explication of intended test interpretations and the collection of empirical evidence supporting these specific interpretations by drawing on several sources of evidence (see *Keppens et al., 2019). Here, the fit between validation measures and intended test use is crucial. Moreover, a sufficient diversity of validation approaches, for example drawing on different theoretical approaches, will help avoiding biases caused by a one-sided view on the construct. Against this background, researchers should also report results that are not in accordance with common theoretical assumptions (e.g., unexpected correlations) to support the further development of theory.

6.5. Limitations

The results of this scoping review are limited by the criteria used to select relevant literature. Owing to the focus on English-language articles published in peer-reviewed journals, books, and book chapters, publications in other languages were not included. Similarly, we omitted publications that aimed at comparable situation-specific competencies without using the terms "noticing" or "professional vision." This concerns instruments such as video-based tests of classroom management expertise (König & Kramer, 2016) and teachers' knowledge of teaching mathematics (Kersting, 2008). Furthermore, recent developments such as the "Video Assessment of Teacher Knowledge" (Wiens, Beck, &

Lunsmann, 2020) are not considered; however, this instrument focuses on knowledge rather than noticing. Nonetheless, this review may help raise awareness of the differentiation between video-based measures of teacher noticing and video-based measures of teacher knowledge that do not rely on teacher noticing.

Finally, the comparably broad range of kappa values suggests that, for some concepts, it was difficult to find a shared understanding. We addressed this issue by presenting tables that included all information relevant for coding, thus ensuring the transparency of our analysis.

6.6. Conclusion

We consider the results of this scoping review to be encouraging with respect to the development of quality test instruments to measure teacher noticing. Several high-quality test instruments were identified as providing measurement approaches and validation strategies that other researchers can draw on. Given the heterogeneity of existing instruments outlined above, the overview provided by this review might support researchers in carefully conceptualizing the noticing construct they aim to measure—including the definitions and nomenclature of the noticing facets addressed—and in judiciously selecting adequate measurement approaches and appropriate validation strategies.

Funding

This work was supported by the German Ministry of Education and Research [grant numbers: 01PK19006A, 01PK19006B].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tate.2022.103970.

References

References marked with an asterisk indicate publications included in the literature review.

Amador, J., Bragelman, J., & Castro Superfine, A. (2021). Prospective teachers noticing: A literature review of methodological approaches to support and analyze noticing. *Teaching and Teacher Education, 99*, Article 103256. https://doi.org/10.1016/j.tate.2020.103256

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). In *Standards for educational and psychological testing*. American Educational Research Association.

Berliner, D. C. (1988). *The development of expertise in pedagogy*. American Association of Colleges for Teachers.

Blomberg, G., Stürmer, K., & Seidel, T. (2011). How pre-service teachers observe teaching on video: Effects of viewers' teaching subjects and the subject of the video. *Teaching and Teacher Education, 27*(7), 1131–1140. https://doi.org/10.1016/j.tate.2011.04.008

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). 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. (2015). 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

Choy, B. H., & Dindyal, J. (2020). Teacher noticing, mathematics. In M. A. Peters (Ed.), *Encyclopedia of teacher education (living)*. Springer. https://doi.org/10.1007/978-981-13-1179-6_241-1.

Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., et al. (2014). Scoping reviews: Time for clarity in definition, methods, and reporting. *Journal of Clinical Epidemiology, 67*(12), 1291–1294. https://doi.org/10.1016/j.jclinepi.2014.03.013

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/H0040957

*Dalvi, T., & Wendell, K. (2017). Using student video cases to assess pre-service elementary teachers' engineering teaching responsiveness. *Research in Science Education, 47*(5), 1101–1125. https://doi.org/10.1007/s11165-016-9547-5

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE.

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

*Dreher, A., & Kuntze, S. (2015a). Teachers facing the dilemma of multiple representations being aid and obstacle for learning: Evaluations of tasks and theme-specific noticing. *Journal für Mathematik-Didaktik, 36*(1), 23–44. https://doi.org/10.1007/s13138-014-0068-3

*Dreher, A., & Kuntze, S. (2015b). Teachers' professional knowledge and noticing: The case of multiple representations in the mathematics classroom. *Educational Studies in Mathematics, 88*(1), 89–114. https://doi.org/10.1007/s10649-014-9577-8

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.

*Fisher, M. H., Thomas, J., Jong, C., Schack, E. O., & Dueber, D. (2019). Comparing preservice teachers' professional noticing skills in elementary mathematics classrooms. *School Science & Mathematics, 119*(3), 142–149. https://doi.org/10.1111/ssm.12324

*Fisher, M. H., Thomas, J., Schack, E. O., Jong, C., & Tassell, J. (2018). Noticing numeracy now! Examining changes in preservice teachers' noticing, knowledge, and attitudes. *Mathematics Education Research Journal, 30*(2), 209–232. https://doi.org/10.1007/s13394-017-0228-0

*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.

Grub, A.-S., Biermann, A., & Brünken, R. (2020). Process-based measurement of professional vision of (prospective) teachers in the field of classroom management: A systematic review. *Journal for Educational Research Online, 12*(3), 75–102.

Huang, Y., Miller, K. F., Cortina, K. S., & Richter, D. (2021). *Teachers' professional vision in action: Comparing expert and novice teacher's real-life eye movements in the classroom*. Zeitschrift für Pädagogische Psychologie, advance online publication. https://doi.org/10.1024/1010-0652/a000313

Jacobs, V. R. (2017). Complexities in measuring teacher noticing: Commentary. In E. O. Schack, M. H. Fisher, & J. A. Wilhelm (Eds.), *Teacher noticing: Bridging and broadening perspectives, contexts, and frameworks* (pp. 273–279). Springer. https://doi.org/10.1007/978-3-319-46753-5_16.

*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

*Johnson, H. L., Dunlap, J. C., Verma, G., McClintock, E., DeBay, D. J., & Bourdeaux, B. (2019). Video-based teaching playgrounds: Designing online learning opportunities to foster professional noticing of teaching practices. *TechTrends, 63*(2), 160–169. https://doi.org/10.1007/s11528-018-0286-5

*Kaendler, C., Wiedmann, M., Leuders, T., Rummel, N., & Spada, H. (2016). Monitoring student interaction during collaborative learning: Design and evaluation of a Training program for pre-service teachers. *Psychology Learning and Teaching, 15*(1), 44–64. https://doi.org/10.1177/1475725716638010

*Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M., & Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers—cognitive versus situated approaches. *Educational Studies in Mathematics, 94*(2), 161–182. https://doi.org/10.1007/s10649-016-9713-8

*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

Kaiser, G., & König, J. (2019). Competence measurement in (mathematics) teacher education and beyond: Implications for policy. *Higher Education Policy, 32*(4), 597–615. https://doi.org/10.1057/s41307-019-00139-z

*Keppens, K., Consuegra, E., Goossens, M., Maeyer, S. de, & Vanderlinde, R. (2019). Measuring pre-service teachers' professional vision of inclusive classrooms: A video-based comparative judgement instrument. *Teaching and Teacher Education, 78*, 1–14. https://doi.org/10.1016/j.tate.2018.10.007

Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement, 68*(5), 845–861. https://doi.org/10.1177/0013164407313369

*Kleinknecht, M., & Gröschner, A. (2016). Fostering preservice teachers' noticing with structured video feedback: Results of an online- and video-based intervention study. *Teaching and Teacher Education, 59*, 45–56. https://doi.org/10.1016/j.tate.2016.05.020

*König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education, 38*, 76–88. https://doi.org/10.1016/j.tate.2013.11.004

König, J., & Kramer, C. (2016). Teacher professional knowledge and classroom management: On the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM-Mathematics Education, 48*(1–2), 139–151. https://doi.org/10.1007/s11858-015-0705-4

König, J., Santagata, R., Scheiner, T., Adleff, A.-K., Yang, X., & Kaiser, G. (2022). Teacher noticing and teacher professional vision: A systematic literature review on conceptualizations and research designs. *Educational Research Review, 36*. https://doi.org/10.1016/j.edurev.2022.100453

Kosko, K. W., Heisler, J., & Gandolfi, E. (2022). Using 360-degree video to explore teachers' professional noticing. *Computers & Education, 180*, Article 104443. https://doi.org/10.1016/j.compedu.2022.104443

Kramer, M., Förtsch, C., Stürmer, J., Förtsch, S., Seidel, T., & Neuhaus, B. J. (2020). Measuring biology teachers' professional vision: Development and validation of a video-based assessment tool. *Cogent Education, 7*(1), Article 1823155. https://doi.org/10.1080/2331186X.2020.1823155

Lachner, A., Jarodzka, H., & Nückles, M. (2016). What makes an expert teacher? Investigating teachers' professional vision and discourse abilities. *Instructional Science, 44*(3), 197–203. https://doi.org/10.1007/s11251-016-9376-y

Mason, J. (2002). *Researching your own practice: The discipline of noticing*. Routledge.

Mason, J. (2011). Noticing: Roots and branches. In M. G. Sherin, V. R. Jacobs, & R. A. Philipp (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 35–50). Routledge.

Mason, J. (2021). Learning about noticing, by, and through, noticing. *ZDM-Mathematics Education, 53*(1), 231–243. https://doi.org/10.1007/s11858-020-01192-4

*Meschede, N., Fiebranz, A., Möller, K., & Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: On its relation and differences between pre-service and in-service teachers. *Teaching and Teacher Education, 66*, 158–170. https://doi.org/10.1016/j.tate.2017.04.010

*Meschede, N., Steffensky, M., Wolters, M., & Möller, K. (2015). Professionelle Wahrnehmung der Lernunterstützung im naturwissenschaftlichen Grundschulunterricht: Theoretische Beschreibung und empirische Erfassung [Professional vision on science lessons in primary school – conceptualization and measurement]. *Unterrichtswissenschaft, 43*(4), 317–335.

*Michalsky, T. (2014). Developing the SRL-PV assessment scheme: Preservice teachers' professional vision for teaching self-regulated learning. *Studies In Educational Evaluation, 43*, 214–229. https://doi.org/10.1016/j.stueduc.2014.05.003

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology, 18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

Nickerson, S. D., Lamb, L., & LaRochelle, R. (2017). Challenges in measuring secondary mathematics teachers' professional noticing of students' mathematical thinking. In E. O. Schack, M. H. Fisher, & J. A. Wilhelm (Eds.), *Teacher noticing: Bridging and broadening perspectives, contexts, and frameworks* (pp. 381–398). Springer International Publishing. https://doi.org/10.1007/978-3-319-46753-5_22.

Noordink, T., Verharen, L., Schalk, R., van Eck, M., & van Regenmortel, T. (2021). Measuring instruments for empowerment in social work: A scoping review. *British Journal of Social Work, 51*(4), 1482–1508. https://doi.org/10.1093/bjsw/bcab054

*Roose, I., Goossens, M., Vanderlinde, R., Vantieghem, W., & van Avermaet, P. (2018). Measuring professional vision of inclusive classrooms in secondary education through video-based comparative judgement: An expert study. *Studies In Educational Evaluation, 56*, 71–84. https://doi.org/10.1016/j.stueduc.2017.11.007

*Roose, I., Vantieghem, W., Vanderlinde, R., & van Avermaet, P. (2019). Beliefs as filters for comparing inclusive classroom situations. Connecting teachers' beliefs about teaching diverse learners to their noticing of inclusive classroom characteristics in videoclips. *Contemporary Educational Psychology, 56*, 140–151. https://doi.org/10.1016/j.cedpsych.2019.01.002

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *American Educational Research Journal, 28*(1), 63–88.

Santagata, R., König, J., Scheiner, T., Nguyen, H., Adleff, A.-K., Yang, X., et al. (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

*Schack, E. O., Fisher, M. H., Thomas, J. N., Eisenhardt, S., Tassell, J., & Yoder, M. (2013). Prospective elementary school teachers' professional noticing of children's early numeracy. *Journal of Mathematics Teacher Education, 16*(5), 379–397. https://doi.org/10.1007/s10857-013-9240-9

Scheiner, T. (2016). Teacher noticing: Enlightening or blinding? *ZDM-Mathematics Education, 48*(1–2), 227–238. https://doi.org/10.1007/s11858-016-0771-2

*Scherrer, J., & Stein, M. K. (2013). Effects of a coding intervention on what teachers learn to notice during whole-group discussion. *Journal of Mathematics Teacher Education, 16*(2), 105–124. https://doi.org/10.1007/s10857-012-9207-2

Schön, D. A. (1983). *The reflective practitioner. How professionals think in action*. Temple Smith.

Seidel, T., Blomberg, G., & Stürmer, K. (2010). "Observer" – Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. Projekt OBSERVE ["Observer" – validation of a video-based instrument to assess professional vision of teachig]. *Zeitschrift für Padagogik, 56*, 296–306.

*Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal, 51*(4), 739–771. https://doi.org/10.3102/0002831214531321

Sherin, M. G. (2001). Developing a professional vision of classroom events. In T. Wood, B. S. Nelson, & J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 75–93). Erlbaum.

Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (Eds.). (2011a). *Mathematics teacher noticing: Seeing through teachers' eyes*. Routledge.

Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (2011). 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). Routledge.

Sherin, M. G., Russ, R. S., & Colestock, A. A. (2011). Accessing mathematics teachers' in-the-moment noticing. In M. G. Sherin, V. R. Jacobs, & R. A. Philipp (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 79–94). Routledge.

Sherin, M. G., Russ, R. S., Sherin, B. L., & Colestock, A. (2008). Professional vision in action: An exploratory study. *Issues in Teacher Education, 17*(2), 27–46.

Sherin, M. G., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education, 60*(1), 20–37.

*Simpson, A., & Haltiwanger, L. (2017). This is the First Time I've Done This": Exploring secondary prospective mathematics teachers' noticing of students' mathematical thinking. *Journal of Mathematics Teacher Education, 20*(4), 335–355. https://doi.org/10.1007/s10857-016-9352-0

Stahnke, R., Schueler, S., & Roesken-Winter, B. (2016). Teachers' perception, interpretation, and decision-making: A systematic review of empirical mathematics education research. *ZDM-Mathematics Education, 48*(1–2), 1–27. https://doi.org/10.1007/s11858-016-0775-y

*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

*Steffensky, M., Gold, B., Holodynski, M., & Möller, K. (2015). Professional vision of classroom management and learning support in science classrooms—does professional vision differ across general and content-specific classroom interactions? *International Journal of Science and Mathematics Education, 13*(2), 351–368. https://doi.org/10.1007/s10763-014-9607-0

*Stürmer, K., Könings, K. D., & Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. *British Journal of Educational Psychology, 83*(3), 467–483. https://doi.org/10.1111/j.2044-8279.2012.02075.x

*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

*Stürmer, K., & Seidel, T. (2015). Assessing professional vision in teacher candidates. *Zeitschrift für Psychologie, 223*(1), 54–63. https://doi.org/10.1027/2151-2604/a000200

*Stürmer, K., Seidel, T., & Holzberger, D. (2016). Intra-individual differences in developing professional vision: Preservice teachers' changes in the course of an innovative teacher education program. *Instructional Science, 44*(3), 293–309. https://doi.org/10.1007/s11251-016-9373-1

*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

*Theelen, H., Beemt, A., & Brok, P. (2019). Using 360-degree videos in teacher education to improve preservice teachers' professional interpersonal vision. *Journal of Computer Assisted Learning, 35*(5), 582–594. https://doi.org/10.1111/jcal.12361

*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

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). Prisma extension for scoping reviews (prisma-scr): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850

*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

*Wiedmann, M., Kaendler, C., Leuders, T., Spada, H., & Rummel, N. (2019). Measuring teachers' competence to monitor student interaction in collaborative learning settings. *Unterrichtswissenschaft, 47*(2), 177–199. https://doi.org/10.1007/s42010-019-00047-6

Wiens, P. D., Beck, J. S., & Lunsmann, C. J. (2020). Assessing teacher pedagogical knowledge: The video assessment of teacher knowledge (VATK). *Educational Studies, 48*(2), 273–289. https://doi.org/10.1080/03055698.2020.1750350

*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.

*Yang, X., Kaiser, G., König, J., & Blömeke, S. (2019). Professional noticing of mathematics teachers: A comparative study between Germany and China. *International Journal of Science and Mathematics Education, 17*(5), 943–963. https://doi.org/10.1007/s10763-018-9907-x