Teaching and Teacher Education 95 (2020) 103146

Contents lists available at ScienceDirect

Teaching and Teacher Education

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

Between authenticity and cognitive demand: Finding a balance in designing a video-based simulation in the context of mathematics teacher education

Elias Codreanu [a, *], Daniel Sommerhoff [b], Sina Huber [a], Stefan Ufer [b], Tina Seidel [a]

[a] Friedl Schöller Endowed Chair for Educational Psychology, School of Education, Technical University of Munich (TUM), Germany
[b] Mathematics Education, Faculty of Mathematics, Computer Science and Statistics, Ludwig-Maximilians-Universität München (LMU), Germany

HIGHLIGHTS

- Assessment of learning is an important but challenging practice for teachers.
- Simulations can be used to measure and support teachers' assessment skills.
- A video-based simulation for pre-service mathematics teachers was developed.
- Data supports the simulation' authenticity and adequate cognitive demand.
- The developed simulation provides a tool to validly measure assessment skills.

ARTICLE INFO

Article history:
Received 6 May 2019
Received in revised form
6 May 2020
Accepted 26 June 2020
Available online 20 July 2020

Keywords:
Assessment skills
Mathematical argumentation
Simulation
Teacher education
Video-based learning environment

ABSTRACT

A key challenge for teachers is the on-the-fly assessment of student learning. Video-based simulations may provide a tool for measuring assessment skills and a basis for learning environments in teacher education. Based on the framework for teaching practice by Grossman et al. (2009), considerations for designing video-based simulations that balance authenticity and cognitive demand are derived. Results show that participants perceived the developed simulation as authentic, were mostly able to rank students according to their overall mathematical argumentation skills and showed potential for learning in their detailed assessment of students. Thus, results indicate the internal validity of the video-based simulation.

1. Introduction

During everyday classroom practice, teachers face a multitude of on-the-fly assessment situations in which they gather information about the learning prerequisites, processes, and outcomes of their students (Herppich et al., 2018; Praetorius, Berner, Zeinz, Scheunpflug, & Dresel, 2013; Ruiz-Primo & Furtak, 2007; Thiede et al., 2015). Across a variety of educational systems worldwide, these assessment situations arise predominantly during the interactions between teachers and students in class (Birenbaum et al., 2006; Furtak et al., 2016; Kingston & Nash, 2011; Klug, Bruder, Kelava, Spiel, & Schmitz, 2013). However, empirical data underlines that many novice teachers struggle with this high complexity in their first years of teaching (Correa, Martínez-Arbelaiz, & Aberasturi-Apraiz, 2015; Dicke, Elling, Schmeck, & Leutner, 2015; Levin, Hammer, & Coffey, 2009). Better preparation for assessment situations during teacher–student interactions is therefore essential.

In line with these findings, one of the main challenges in successful teacher education is providing learning environments for future teachers that allow them to acquire practice-oriented knowledge (Cochran-Smith & Zeichner, 2005; Grossman & McDonald, 2008). Practice-oriented knowledge is characterized by knowledge structures with strong links between conceptual understanding and specific practice-based applications. Typically,

* Corresponding author. Technical University of Munich (TUM), Arcisstraße 21, 80333, Munich, Germany.
E-mail address: elias.codreanu@tum.de (E. Codreanu).

https://doi.org/10.1016/j.tate.2020.103146

real classroom settings as the initial environment for knowledge acquisition and application are seen as too complex for most novice teachers to link conceptual understanding to practical situations (Stokking, Leenders, Jong, & van Tartwijk, 2003). Therefore, the framework for teaching practice developed by Grossman et al. (2009) proposes strategies to approximate teaching practice by beginning the learning process in an environment low in complexity, increasing from there until the point of applying knowledge to a real classroom setting. While practice in real classrooms represents the highest authenticity and complexity concerning professional tasks, other complexity-reduced practices, such as the analysis of transcripts or vignettes, observation of videos, and participation in role-plays or simulations, are still regarded as authentic.

Over the last decade, the use of classroom video has become an important element in teacher education (Gaudin & Chaliès, 2015; Kang & van Es, 2018). Research has shown that video-based approaches represent authentic and complexity-reduced representations of practice, providing particular opportunities for pre-service teachers to acquire practice-oriented knowledge (Seidel, Stürmer, Blomberg, Kobarg, & Schwindt, 2011; Sherin, Linsenmeier, & van Es, 2009; Tekkumru-Kisa & Stein, 2017; Van Es & Sherin, 2008). However, the possible negative effects of processing video-based information due to suboptimal instructional designs and high cognitive demand are also discussed in the literature (Derry, Sherin, & Sherin, 2014). Therefore, balancing authenticity and cognitive demand remains a major challenge in designing video-based environments (Blomberg, Renkl, Sherin, Borko, & Seidel, 2013).

So far, video in teacher education research has been mainly applied as a medium in digital tools or teacher education courses. However, in other professional fields, simulations play an important role in supporting practice-oriented knowledge acquisition (Chernikova et al., 2019). Therefore, given recent advances in video-based research, it seems timely to start developing and testing simulations for teacher education. In this article, we present a first attempt in this direction. Staged classroom videos as well as instructional design features that allow participants to intervene in a situation and influence the course of action were implemented in the newly created simulation.

The present article provides first insights into the internal validity of this newly developed video-based simulation. The added-value can be viewed in three ways: First, the study provides conceptual value, since it exemplarily outlines how concepts of the framework for teaching practice by Grossman et al. (2009) and recent advances in video-based research can serve as a basis for simulation design. Second, the study provides methodological value, since the content of the simulation is highly driven by evidence stemming from mathematics education in terms of typical student misconceptions related to mathematical argumentation, which was embedded in the instructional design of video-based simulations. Third, the insights provided by this study have practical relevance in international teacher education contexts, since the study itself can serve as a prototype for the design of further simulations.

1.1. Teacher assessment skills: a central professional requirement

Teacher assessment skills are a central element for teachers' practice and a basic requisite for successful teaching and learning in the twenty-first century (Darling-Hammond & Bransford, 2007). In many countries, competency frameworks for teachers include skills related to the assessment of students, with multiple educational reforms having been established around such frameworks (European Commission, 2013). While the assessment of students' learning and thinking processes is one of the topics that are currently receiving focus in academic discussions about the teaching profession (Heitzmann et al., 2019; Loibl, Leuders, & Dörfler, 2020; Südkamp & Praetorius, 2017), experts highlight that there is still considerable potential for the further development of these skills in existing teacher education (Schrader, 2017). Despite the increasing amount of attention paid toward teachers' assessment skills within teacher professionalization, there are so far only a few tools and environments within teacher education that allow for validly measuring and supporting the improvement of these skills (Praetorius, Lipowsky, Karst, Lazarides, & Ittel, 2012; Südkamp, Möller, & Pohlmann, 2008).

The lack of tools and environments in teacher education may be due to the fact that researchers are still exploring the prerequisites and processes involved in on-the-fly formative assessment (Alonzo, 2011; Black & Wiliam, 2009; Furtak et al., 2016), while research has focused on summative assessment situations for a long time, e.g., in grading (Schrader & Helmke, 1987; Spinath, 2005; Südkamp, Kaiser, & Möller, 2012). Moreover, there is a certain disagreement regarding the investigation of teachers' assessment skills: While there is a considerable amount of research on the accuracy of teacher assessments (Helmke & Schrader, 1987; McElvany et al., 2009; Südkamp, Möller, & Pohlmann, 2008), substantial discrepancies between findings collected through different methods make drawing conclusions difficult (Spinath, 2005). For example, some researchers asked teachers to rank students, leading to quite a general assessment based on a comparison of all students within a social setting (McElvany et al., 2009), while others compared teachers' assessments of students to experts' evaluations (Karst & Förster, 2017). Finally, some researchers focused on the evaluation of specific subskills, knowledge aspects, and other characteristics exhibited by students that might serve as cues for the assessment of the skill-to-analyze.

In the context of mathematical argumentation skills, research has highlighted the value of mathematical content knowledge (Weigand et al., 2014), methodological knowledge (Heinze & Reiss, 2003), and the use of problem-solving strategies (Schoenfeld, 1992) as a basis for such skills (Ufer, Heinze, & Reiss, 2008). Accordingly, the accurate assessment of each of these three content dimensions can be regarded as a prerequisite for the formative and summative assessment of students' mathematical argumentation skills. In particular, focusing on these three content dimensions supports pre-service teachers in learning that these content dimensions are likely not represented homogeneously within each student but that students have distinct profiles regarding these dimensions. Next to profiles of overall high or low skills in all three content dimensions, mixed profiles, particularly concerning methodological knowledge and the use of problem-solving strategies, have been described. These student profiles, stemming from evidence in mathematics education (Reiss & Ufer, 2009), can serve as a basis for designing a video-based simulation, e.g., for staging videos and defining roles for the simulated students, representing avatars for identified profiles in research.

1.2. Authenticity of a simulation: an approximation of practice

To support pre-service teachers in acquiring practice-oriented professional knowledge and skills, experts call for a close relation to real-world professional situations (Blömeke, Gustafsson, & Shavelson, 2015; Herppich et al., 2018; Kaiser et al., 2017; Koeppen, Hartig, Klieme, & Leutner, 2008). In common teacher education programs, this is often realized in the form of school internships; however, these can be rather demanding for pre-service teachers and represent quite complex situations (Stokking et al., 2003). For example, even if pre-service teachers in

internships only observe teaching, they often lack sufficient professional knowledge (see also Förtsch et al., 2018) to notice and interpret significant situations in the classroom and thus may miss learning opportunities (Seidel & Stürmer, 2014). Therefore, teacher education must find additional ways to a) determine the current skills of pre-service teachers and b) provide environments in which they can practice relevant and specific skills. Grossman et al. (2009) call such opportunities to engage in practices that are similar to professional teaching practice *approximations of practice*.

Two aspects of such approximations of practice in teacher education should be considered and need to be balanced: authenticity regarding real-world practice and complexity regarding required cognitive demands. Other research approaches, which also refer to the balance between authenticity and cognitive demand, point toward simulations, defined as something that "imitates one process by another process" (Hartmann, 1996, p. 83). Both approaches agree that, from the perspective of the learner, it is essential that the trained behavior can be easily transferred to real-world situations. Thus, the learner should perceive the represented situation as authentic (Seidel, Blomberg, & Stürmer, 2010) and feel sufficiently present (Schubert, Friedmann, & Regenbrecht, 2001) to become involved in the actual learning situation. Moreover, to exhibit their skills in artificial situations, learners must be cognitively involved (Dankbaar et al., 2016) and focus their attention on the situation (Witmer & Singer, 1998). Research shows that the simulation's presented opportunity to engage in the process of action supports the cognitive involvement of learners (Paas, Tuovinen, Tabbers, & van Gerven, 2003). This is also supported by Chen and Wu (2012), emphasizing that in order to learn from approximations of practice, students must be motivated to be active participants in the learning process.

The call for approximations of practice in teacher education is currently often answered by the use of videos (Blömeke, Gustafsson, & Shavelson, 2015; van Es, Tekkumru-Kisa, & Seago, 2020). Videos capture the authenticity of classroom practice, bringing to life the work of teaching for careful study (Brophy, 2004; Kang & van Es, 2018). Previous research has shown the success of using video in investigating and fostering the acquisition of relevant teacher skills, e.g., in mathematics teaching and other domains (Friesen, Kuntze, & Vogel, 2018; Kaiser, Busse, Hoth, König, & Blömeke, 2015). In particular, prior research has shown that classroom situations represented in videos are perceived as authentic (Seidel et al., 2010). In some approaches, instead of videos demonstrating real-world classroom situations, they are used to exhibit pre-planned, artificial classroom situations – then called staged videos. This type of video entails the risk of reduced authenticity regarding its content and representation of real-world situations. In particular, the selection and composition of the video content, as well as the surrounding setting can contribute to their similarity to professional teaching practice. Still, the research underlined that staged classroom videos are generally experienced as authentic by pre-service teachers (Böttcher & Thiel, 2018; Piwowar, Barth, Ophardt, & Thiel, 2017). Furthermore, staged videos provide several benefits in terms of reduced complexity, since they may allow pre-service teachers to more clearly focus on specific events and eliminate other possibly distracting factors such as background noise, peripheral movement, and actions. Therefore, a video-based simulation using staged videos may provide an authentic approximation of practice for pre-service teachers to investigate and practice relevant teaching skills, such as the on-the-fly assessment of student thinking.

1.3. Cognitive demand of a simulation: task difficulty

Although it is widely accepted that a task within an approximation of practice should be as authentic as possible, its difficulty must also be considered. The difficulty of a task is strongly related to the acquired knowledge base of the learner; many difficulties that pre-service teachers face can be attributed to a lacking connection between the complex requirements of practical tasks and the acquired conceptual knowledge (Bauer & Prenzel, 2012; Grossman & McDonald, 2008). In particular, it appears to be difficult for students to relate conceptual professional knowledge acquired at university to specific practical assessment situations in classrooms (Alles, Apel, Seidel, & Stürmer, 2019). Therefore, when designing tasks for simulations, a balance between authenticity and the level of cognitive demand should be considered, allowing participants to link the practical task to their acquired conceptual knowledge. This concept, called the *decomposition of practice*, involves breaking down practice into its constituent parts for the purposes of teaching and learning (Grossman et al., 2009).

One way to achieve the decomposition of practice is to break down a task within a simulation to a level that pre-service teachers can master individually. When developing assessment tools for teacher education, this level should ideally be designed in a way that allows most participants to be grouped around a medium level of difficulty with enough potential for higher levels of difficulty, which allows for the measurement of positive learning developments (cf. Seidel & Stürmer, 2014). Besides, task difficulties can be varied in meaningful ways by, e.g., starting with tasks that are likely to be easier before moving on to more cognitively demanding tasks. When applying this perspective to a simulation that measures pre-service teachers' skills in assessing students' mathematical argumentation skills, a measure of its cognitive demand could involve comparing participants' judgments of observed student skills with expert judgments, thus using a measure of judgment accuracy (Südkamp & Möller, 2009). In addition, an easier task could involve ranking the observed students according to their overall mathematical argumentation skills. A detailed assessment of students' mathematical argumentation skills considering multiple underlying content dimensions, such as mathematical content knowledge or methodological knowledge, would likely be more difficult to provide. Therefore, the selection of assessment tasks with varying levels of difficulty is important to address the respective cognitive demand of all participants.

1.4. Aim of study

The main aim of the present study was to provide first insights into the internal validity of a newly developed video-based simulation by balancing authenticity and cognitive demand. In particular, the video-based simulation was intended to be applicable in initial teacher education and useable for measuring pre-service teachers' acquired assessment skills at various times throughout a teacher education program, thus allowing for adequate room for increased performance. The first goal of this study was to investigate whether participants perceived the video-based simulation as an authentic approximation of practice, thus allowing for a contextualized, valid method for measuring assessment skills. The second goal of the study was to investigate if the decomposition of the professional situation resulted in an adequate level of cognitive demand, measured by participants' success in the assessment tasks and, therefore, the ability to measure a range of pre-service teachers' assessment skills.

RQ 1: How do pre-service teachers experience the video-based simulation regarding an authentic approximation of practice?

Participants' perceptions of a high level of authenticity in the simulations' videos, the embedding of the videos in the simulation,

and the participants' reported cognitive involvement and motivation were anticipated to provide empirical support for a valid approximation of practice.

RQ 2: How do the two tasks in the video-based simulation indicate adequate levels of difficulty and distributions of student performances in mastering the assessment situation?

(i) Task 1: Provide a correct ranking of the represented students according to their overall mathematical argumentation skills.

(ii) Task 2: Produce accurate, detailed assessments of the specific skills and characteristics of mathematical argumentation skills as exhibited by the students in the video-based simulation.

Based on the observed distribution of participants' responses, adequate levels of difficulty were given if participants reached an average medium score. This implies that the measurement of participants' assessment skills via the two tasks exhibits neither ceiling nor floor effects. Nevertheless, the scoring distribution of the participants' responses for each task should show enough variance to measure differences in their assessment skills. The ranking of students according to their overall mathematical argumentation skills (*Task 1*) was hypothesized to be less difficult than a detailed assessment of the specific aspects of the students' argumentation skills (*Task 2*).

2. Method

2.1. Participants

The recruitment of participants took place within a course in a German bachelor's program for pre-service high school teachers. The pre-service teachers participated voluntarily after a regular meeting for the course. Participation was remunerated but did not influence the course's grading nor the students' passing. In the design and implementation of the study, researchers adhered to the standards of the Ethical Principles of Psychologists and Code of Conduct (American Psychological Association, 2017). The sample consisted of 28 pre-service high school teachers (13 female, 15 male). The average age was 21.5 ($SD = 2.0$). The course, from which the data collection took place was for pre-service high school teachers in their third semester of their bachelor's program. As in this program, the pre-service high school teachers go through the same curriculum, most of the participants in the study were in their third semester. Most participants had completed two university mathematics modules (each module includes two lectures and two exercises per week) in previous semesters. Likewise, most participants had completed two modules focusing on mathematics education and two modules in pedagogy and educational psychology. All participants had gained at least some practical experience in teaching. Additionally, most had completed about two weeks of internships in schools. During this time, they had taught, on average, 2.6 ($SD = 1.3$) lessons on their own. Besides, all participants had experience in private tuition to students in mathematics ($M = 2.9$ years; $SD = 1.3$).

2.2. Design of the video-based simulation

To design a video-based simulation to investigate pre-service teachers' assessment skills, the research team developed a set of staged video clips depicted typical one-on-one teacher–student interactions during class, which were then embedded in a simulation based on the principles of problem-oriented learning (Gräsel, Fischer, & Mandl, 2000) and student scaffolding (Tabak & Kyza, 2018; see also the recent meta-analysis on scaffolding in simulations by; Chernikova et al., 2019). Each video clip includes a simulated teacher and one of four simulated students from seventh grade at a German high school (Andreas, Barbara, Christian, and Doris). In the video clips, a simulated teacher and the simulated students are discussing the students' progress and argumentation in the context of a geometry proof task: The simulated students must prove that opposite sides of a parallelogram are of equal length based on the information that parallelogram pairs of sides are parallel. The participants of the simulation take on the role of a pre-service intern – matching their prior experiences – asked to assess the mathematical argumentation skills of these four simulated seventh graders during the course of the simulation.

For the production of the staged video clips, the research team followed Dieker et al.'s (2009) recommendations, which contain three consecutive phases: first, the selection of evidence-based practices; second, vignette script development; and third, video production. During the first phase, the team conducted a literature review of important mathematical student skills and their assessment in school. Based on the review, the team identified mathematical argumentation skills as a relevant content for the assessment situation. Studies from mathematics education have identified different mathematical content dimensions as predictive of students' performance in handling mathematical argumentations in proofs (Sommerhoff, Ufer, & Kollar, 2015; Schoenfeld, 1992). Based on Ufer, Heinze, & Reiss (2008) and Sommerhoff, Ufer, & Kollar (2015), the current study focused on three of these dimensions: students' mathematical content knowledge (Weigand et al., 2014), methodological knowledge (Heinze & Reiss, 2003), and problem-solving strategies (Schoenfeld, 1992). During the second phase, the research team developed a contextual frame and detailed scripts for the staged video clips of the four simulated students. Guided by van Hiele's model for describing the development of the geometric thinking of children (Usiskin, 1982), four different student profiles were designed: Andreas (profile A), Barbara (profile B), Christian (profile C), and Doris (profile D).[1] The profiles represented different mathematical argumentation skill levels, varying on the three selected mathematical content dimensions. Andreas (profile A) was designed to have the weakest mathematical argumentation skills, Doris (profile D), the strongest, and Barbara (profile B) and Christian (profile C) medium-low and medium-high mathematical argumentation skill levels, respectively. A script was written for each student profile, containing the detailed verbal exchanges of the teacher–student interactions as well as blueprints of the simulated students' sketches and records in the exercise books. During the third phase, the research team created and edited video footage to generate a representation of teaching practice. The footage contained eight staged video clips per student profile, with consecutive teacher–student interactions regarding the geometry proof task. The teacher–student interactions included the progress, thoughts, and questions each simulated student shared during their short discussions with the simulated teacher. One camera additionally showed the sketches and records of the simulated student whenever they became relevant in the conversation (see Fig. 1). Each video clip lasted approximately 1 min ($M = 71$ s, $SD = 22$), and each included cues related to at least two of the three mathematical content dimensions. Besides the verbal statements of the simulated students, cues related to these content dimensions could also be found in the students' sketches and records.

[1] The names of the student profiles have been changed for this article to increase readability. In the original simulation, the student profiles had arbitrary German first names.

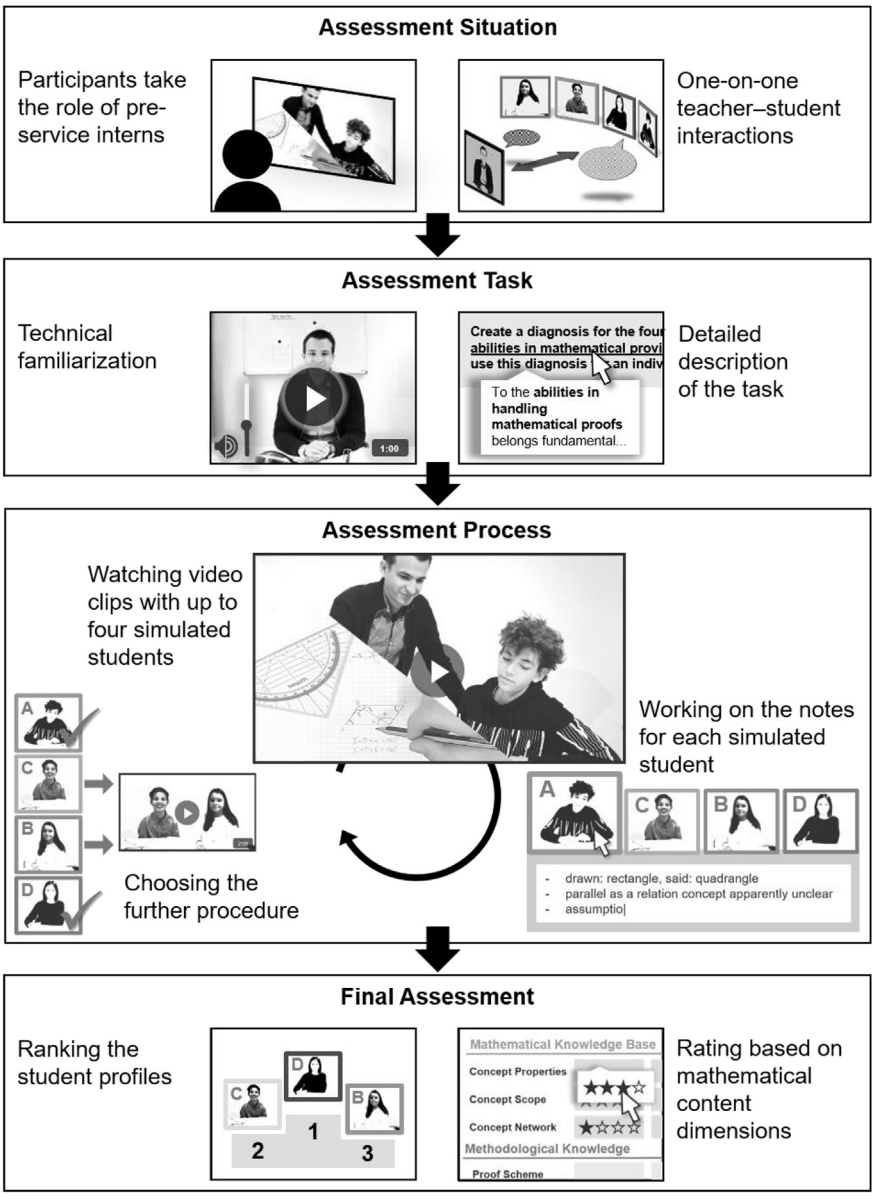

Fig. 1. Design of the video-based simulation.

The staged video clips were then embedded in the simulation. The video-based simulation had an underlying structure consisting of four main parts (see Fig. 1). In part one, the simulation started by familiarizing the participants with the assessment situation of the simulation: Participants were welcomed to the simulated teacher's classroom as observing pre-service interns. Consequently, the participants were introduced to their assessment task: The simulated teacher asked them to assess the mathematical argumentation skills of four simulated students so that he could choose tasks for their individual learning support in subsequent classes. Participants received information on the assessment purpose and the three content dimensions of mathematical argumentation skills on which they were to focus. In part three, the assessment situation, participants could then work independently in the simulated classroom situation to gather information about the simulated students by watching the video clips and taking notes. Initially, the participants observed each simulated student once in their interaction with the simulated teacher. After this first round which included the watching of four video clips and every subsequent round of video watching, participants were able to interact with the simulation and determine the future process by choosing to either observe the simulated student further or deciding they had gathered sufficient information for the assessment of this particular simulated student. In total, a maximum of 20 video clips could be observed, and participants were free to decide how to distribute these observations among the simulated students. This limitation of 20 video clips was set to simulate an authentic classroom situation in which teachers must divide their attention between all students in a class instead of having an unlimited number of interactions with one (or four) specific students. In addition, this limitation was also set to keep participation in the simulation feasible and comparable across the participants' assessment process. The notes participants were allowed to take for each observation during the assessment process would help during the fourth part of the simulation. Here, participants were asked to formulate their notes into a written assessment of each simulated student,

which was introduced in the simulation as to be passed on to the simulated teacher. Then, participants ranked the simulated students according to their overall mathematical argumentation skills from weakest to strongest (*Task 1*). Subsequently, the participants assessed the students' mathematical argumentation skills regarding the three mathematical content dimensions using Likert response items (*Task 2*).

2.3. Measures and instruments

2.3.1. Authenticity of simulation

To assess the extent to which the video-based simulation was perceived as an authentic approximation of practice, participants rated their perception on four different scales after working with the simulation (see Table 1). The two scales for perceived authenticity consisted of three items each and were both adapted from Schubert et al. (2001). They were used to address the authenticity of the videos showing the teacher–student interactions (authenticity of videos) and the representation of the assessment situation in the entire simulation (authenticity of simulation). Additionally, two scales with four items each and addressing the cognitive involvement (Vorderer et al., 2004) and motivation of the participants were implemented. Participants rated each of the four scales on a five-point Likert response scale. The scale for motivation is based on the expectancy-value theory (Wigfield & Eccles, 2000) and measures the value factor of motivation by asking about the value of the assessment task.

2.3.2. Cognitive demand of simulation

The difficulty of the assessment tasks in the video-based simulation, which served as indicators of its adequate demand for participants, was determined using two approaches (Helmke & Schrader, 1987; McElvany et al., 2009): (i) participants' assessment of the students' overall mathematical argumentation skills in rank order and (ii) their detailed assessment of each student regarding the three content dimensions.

Rank order assessment of student profiles. By design, the student profiles could be ranked, as the students exhibited mostly coherent cues related to their mathematical argumentation skills in the video clips, which highlighted their different skill levels. For a correct ranking assessment of the profiles as "A B C D" (from low to high), participants received three points. For each switching of two profiles in their ranking, one point was subtracted, creating a range from zero to three points (e.g., "A B D C" switched C and D, resulting in a score of 2; "A D B C" switched B and D, and C and D, resulting in a score of 1).

Accuracy of detailed assessment: student profiles and content dimensions. The participants assessed each of the four simulated students according to three different mathematical content dimensions: mathematical content knowledge, methodological knowledge, and problem-solving strategies (see Table 2). For the two knowledge dimensions, participants assessed the four simulated students on three items each. For the problem-solving strategies dimension, participants answered two items for each simulated student. Each of the items addressed a distinct aspect of the students' mathematical argumentation skills in terms of geometry proofs. The response format for these items was a four-point Likert response scale.

Prior to the study, two mathematics education researchers independently rated the four student profiles regarding each of the above-mentioned aspects with a substantial interrater agreement (Cohen's $\kappa = 0.80$). A consensus approach was applied in case of differences. The accuracy of the participants' ratings was based on this solution. For each item, participants received one point in the case of agreement with this solution and zero points for non-agreement. Assessment accuracy for each specific student profile

Table 1
Description of the scales for authenticity: Number of items (N), items, and internal consistency (Cronbach's Alpha).

Scales	*N*	Items	α
Authenticity of videos	3	I think the video clips are authentic. The video clips came across like a real-life situation. The experience of the video clips was similar to the experience in a real-life situation.	.74
Authenticity of simulation	3	I think the simulation is authentic. The tasks I had to cope during the simulation came across like a real-life situation. The experience in the simulation was similar to the experience in a real-life situation.	.71
Cognitive involvement	4	I focused heavily on the situation. In the meantime, I have forgotten that I am participating in a study. I have immersed myself in the situation. I was fully committed.	.76
Motivation	4	I think it is important to be able to cope with these tasks. Even if these tasks are not part of a graded assignment, it is important to succeed in them. It is useful to deal with these tasks. It would be useful to deal with these tasks, as it is generally useful to be able to cope with these types of tasks.	.85

Note. Likert response scale from 0 (disagree) to 4 (agree).

Table 2
Description of the mathematical content dimensions: Number of items (N) and items.

Scales	*N*	Items
Mathematical content knowledge	3	The student knows the characteristics of basic terms in geometry. The student has prototypical conceptions for basic terms in geometry. The student knows relationships and connections between basic terms in geometry.
Methodological knowledge	3	The student knows what kind of arguments are valid in a proof. The student knows that a proof begins with the conditions and ends with the claim. The student knows that a proof is an entire sequence of arguments.
Problem-solving strategies	2	The student can use different heuristic strategies to solve problems independently. The student can plan his or her course of action independently, monitor it, and adjust it if necessary.

Note. Likert response scale from 1 (disagree) to 4 (agree).

and mathematical content dimension was defined as the mean score of the items of the particular content dimension for that student profile. The overall assessment accuracy of a student profile or content dimension was calculated as the mean of the relevant items.

2.4. Formative expert feedback on the simulation

Additionally, the research team used an external validation to increase the quality of the simulation. In different phases of the development of the video-based simulation, the research team incorporated the opinions of external mathematics education experts (criteria: more than five years of secondary school experience and currently active in teacher education). In an initial feedback phase, researchers interviewed three external mathematics education experts after reviewing the entire prototype of the video-based simulation. Their feedback on the instructions and handling of the simulation influenced changes made to its design (e.g., which information about the classroom situation was necessary and how and when notes could be taken during the observation process). In a second feedback phase, seven external mathematics education experts worked on the video-based simulation and were asked to evaluate it regarding a) the authenticity of the simulated situation, b) the assessment tasks, and c) the measurement of the assessment skills of the participants. As a result, the authenticity of the simulated situation presented in the video clips ($M = 2.81$, $SD = 0.88$) and the authenticity of the simulation in terms of its instructions ($M = 2.14$, $SD = 0.79$) were rated generally positively by these experts (see above for a description of the scales). In total, the experts' ratings, as well as their open comments, highlighted that they perceived the video-based simulation as an authentic approximation of an assessment situation. The experts pointed out that the task was instructed in a precise and understandable way. They considered the student profiles to be coherently designed with explicit cues for the three content dimensions. The comments from these external reviews of the video-based simulation indicate that the student profiles represented mathematical argumentation skills as described in the literature. Regarding the measurement of the assessment skills, the expert reviews stated that the items for the detailed assessment were suitable for measuring the pre-service teachers' skills in this situation. Their reviews from the second feedback phase showed that these external mathematics education experts were largely in agreement with the solution for the detailed assessment developed by the two mathematics education researchers.

2.5. Analyses

Descriptive statistics were computed to describe and explore participants' perception of the simulation as an approximation of practice. Descriptive statistics are presented in this article for the perceived authenticity of the videos and the simulation as well as the participants' cognitive involvement and motivation. To explore the difficulty of the assessment tasks, descriptive statistics for the difficulty of the rank order assessment and the accuracy of the detailed assessment were computed. Additionally, one-way repeated measures analyses of variances (rmANOVAs) of the accuracy of the detailed assessment were computed to reveal differences in accuracy between the assessments of the student profiles and content dimensions. All participants were included in the analyses, as no specific outliers (e.g., not regarding superficial processing of the simulation) could be identified. For testing the conjectures empirically, the rmANOVAs were computed with contrasts and bonferroni-adjusted post-hoc analyses for multiple comparisons to analyze any differences in the accuracy of the student profile assessments or content dimensions in more detail.

As a prerequisite for analyzing the differences in the accuracy of the four student profiles, the Shapiro-Wilk test was computed. It showed the normality of the accuracy in rating profile B (Barbara) ($p = .254$) and C (Christian) ($p = .097$)[2], however, the test yielded significant results for profile A (Andreas) ($p = .002$) and D (Doris) ($p = .035$). Due to the robustness of a rmANOVA under an application of non-normally distributed data (Schmider, Ziegler, Danay, Beyer, & Bühner, 2010), the rmANOVA was computed with no further adjustments as the Mauchly test for violations of sphericity was not significant ($p = .606$). Regarding differences in the accuracy of the three content dimensions, the Shapiro-Wilk test revealed the normality of the assessment accuracy for mathematical content knowledge ($p = .842$) and methodological knowledge ($p = .319$). Even though the test yielded a significant deviation from normality for problem-solving strategies ($p = .007$), the rmANOVA was computed with no further adjustments as the Mauchly test for violations of sphericity was not significant ($p = .235$).

3. Results

3.1. Descriptive results for participants' use of the simulation

In the present study, the median working time of the participants was 1.2 h for the video-based simulation. Most participants watched 13 video clips ($M = 12.68$, $SD = 5.64$), but the full range of the allowed number of video clips (4–20) was used across participants. The participants decided to watch approximately the same number of video clips across the different student profiles. They watched an average of 3.04 ($SD = 1.60$) video clips for the weaker student profile Andreas, 3.36 ($SD = 1.62$) video clips for the medium-low student profile Barbara, 3.07 ($SD = 1.46$) video clips for the medium-high student profile Christian, and 3.21 ($SD = 1.71$) video clips for the high-performing student profile Doris. For all student profiles, the participants used the provided text boxes for taking notes during the assessment situation. In the last part of the video-based simulation, all participants gave their final assessment. It took them approximately half a minute ($Mdn = 30$ s) to rank the student profiles according to their overall mathematical argumentation skills. In about a minute per student profile, they answered the questions for the detailed assessment (profile A: $Mdn = 63$ s, profile B: $Mdn = 77$ s, profile D: $Mdn = 53$ s). Only for Christian (profile C) did participants use approximately 2 min ($Mdn = 121$ s). This might be due to the fact that this student profile was the first profile shown to participants during the simulation, and the questions were still new for the participants.

3.2. Authenticity of simulation

On average, the participants evaluated the authenticity of the videos and simulation, as well as their cognitive involvement and motivation in the upper half of the Likert response scale (see Table 3). Their motivation was rated highest, with an average of 3.25 in the top fourth of the response scale, followed by the authenticity of the videos with an average of 2.73 in the top third of the response scale. The standard deviations for all the scales were to be similarly small ($SD < 0.81$). None of the participants rated on the lowest end of the response scale. Between 50% (authenticity of simulation) and 93% (motivation) of the given answers were in the highest two ratings on the response scales. In contrast, only 4% (one participant; for both authenticity scales and motivation) to 18%

[2] The following significance levels are used throughout the article: $^{***}p < .001$, $^{**}p < .010$, $^{*}p < .050$.

Table 3
Descriptive statistics for authenticity: Percentage of choices, mean values (M), and the standard deviation (SD) of the scales.

Scales	Does not apply %	Hardly applies %	Partly applies %	Mostly applies %	Applies %	*M*	*SD*
Authenticity of videos	0	4	32	54	11	2.73	0.70
Authenticity of simulation	0	4	46	39	11	2.50	0.75
Cognitive involvement	0	18	21	54	7	2.37	0.81
Motivation	0	4	4	43	50	3.25	0.70

Note. $N = 28$. Likert response scale between 0 (does not apply) and 4 (applies).

(cognitive involvement) of the answers were in the lowest two ratings on the response scales.

3.3. Cognitive demand of simulation

3.3.1. Rank order assessment

Overall, participants received an average score of 2.57 points ($SD = 0.57$) (in a range from zero to three points) on the rank order assessment task. Four different rank orders were observed for the sample (see Table 4). Regarding the difficulty of this task for the participants, 61% of the participants correctly answered the ranking task by arranging the student profiles from A to D. Thirty-six percent of the participants switched the order of two student profiles: either the two medium-level students (B and C) or the two stronger students (C and D). One participant switched three student profiles.

Table 4
Descriptive statistics for the rank order of the student profiles: Number of participants (N), percentage of participants (%), and scoring (Score).

Rank order	N	%	Score
A B C D	17	61	3
A C B D	7	25	2
A B D C	3	11	2
C A B D	1	3	1

Note. $N = 28$. The score ranges from 0 to 3.

3.3.2. Accuracy of detailed assessment

Table 5 shows the accuracy of the participants' detailed assessment regarding the three content dimensions (agreement with expert judgments) for the four student profiles. Descriptive data show that the assessment accuracy was highest for Doris, the strongest student (M = 0.68, SD = 0.21). Assessment accuracy was slightly lower for the weakest student, Andreas (M = 0.54, SD = 0.32), and lowest for the medium-level students, Christian (M = 0.39, SD = 0.22) and Barbara (M = 0.37, SD = 0.18). The assessment accuracy of the content dimensions for the student profiles was highest for the items belonging to the mathematical content knowledge of the students (M = 0.60, SD = 0.22), lower for methodological knowledge (M = 0.46, SD = 0.14), and the lowest for problem-solving strategies (M = 0.42, SD = 0.15). The overall assessment accuracy average in this video-based simulation was 0.50 (SD = 0.14). In other words, approximately half of the participants' ratings matched the solution.

3.3.3. Differences in the accuracy of the detailed assessments

The first rmANOVA showed significant overall differences in the accuracy of assessing different student profiles ($F(3, 81) = 12.36$, $p < .001$), highlighting a large effect size of $\eta^2 = 0.23$. Planned contrasts revealed a statistically significant difference in the accuracy of assessing the two outmost student profiles, the weakest and strongest (Andreas and Doris), and the two medium-level student profiles (Barbara and Christian). There was a mean difference of 0.23 between both types of profile (SE = 0.04, Bonferroni-adjusted $p < .001$; see Fig. 2). In addition, post-hoc pairwise comparisons revealed a significant difference in the accuracy of assessing the strongest student, Doris ($M = 0.68$, $SD = 0.21$), and the two medium-level students. The difference in the accuracy of assessing the profile Doris and the profile Barbara ($M = 0.37$, $SD = 0.18$) was 0.31 ($SE = 0.06$, Bonferroni-adjusted $p < .001$) and the difference in the accuracy of assessing the profile Doris and the profile Christian (M = 0.39, SD = 0.22) was 0.29 (SE = 0.06, Bonferroni-adjusted $p < .001$). A significant difference in the accuracy of the student profiles for Andreas (M = 0.54, SD = 0.32) and Barbara (M = 0.37,

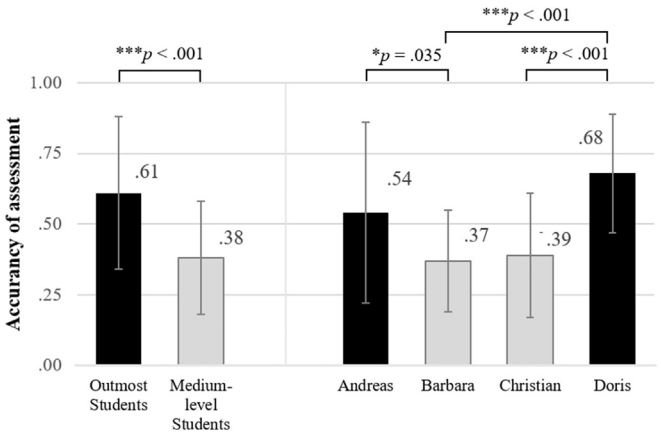

Fig. 2. Detailed assessment accuracy for all four simulated student profiles. Left: comparison of the assessment accuracy of the two outmost and the two medium-level student profiles. Right: comparison of assessment accuracy of the four individual student profiles.

Table 5
Descriptive statistics for level of difficulty: Mean values (M) and standard deviation (SD) of accuracy for each student profile and each content dimension.

Scales	Andreas	Barbara	Christian	Doris	Total
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Mathematical content knowledge	.61 (0.37)	.44 (0.33)	.48 (0.40)	.83 (0.28)	.60 (0.22)
Methodological knowledge	.57 (0.45)	.44 (0.22)	.36 (0.24)	.56 (0.27)	.46 (0.14)
Problem-solving strategies	.38 (0.22)	.29 (0.25)	.36 (0.33)	.64 (0.36)	.42 (0.15)
Total	.54 (0.32)	.37 (0.18)	.39 (0.22)	.68 (0.21)	.50 (0.14)

Note. $N = 28$. The accuracy of the assessment is expressed as the normalized agreement with the solution and ranges from 0.00 (no agreement with the solution) to 1.00 (perfect agreement with the solution).

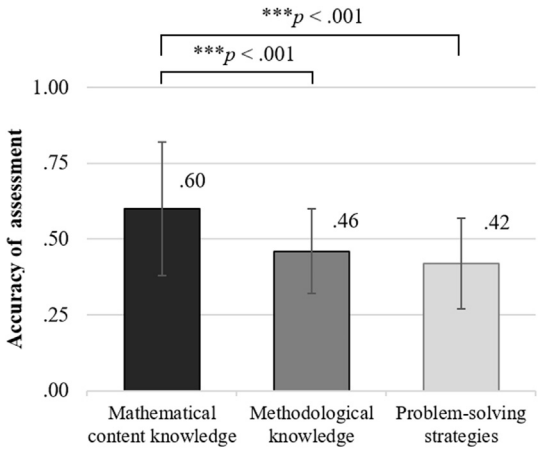

Fig. 3. Detailed assessment accuracy for all content dimensions.

SD = 0.18) was found: 0.17 (SE = 0.06, Bonferroni-adjusted p = .035) (see Fig. 2).

The second rmANOVA determined that the mean accuracy of the participants revealed statistically significant overall differences between the three mathematical content dimensions ($F(2, 54) = 16.40$, $p < .001$) with a large effect size of $\eta^2 = 0.17$. Post-hoc analyses revealed significant differences in the accuracy of the mathematical content knowledge ($M = 0.60$, $SD = 0.22$) and the methodological knowledge ($M = 0.46$, $SD = 0.14$) assessments of the students, with a difference of 0.14 ($SE = 0.03$, Bonferroni-adjusted $p < .001$), as well as their problem-solving strategies (M = 0.42, SD = 0.15), with a difference of 0.19 (SE = 0.04, Bonferroni-adjusted $p < .001$) (see Fig. 3).

4. Discussion

This article focuses on the use of video-based simulations as assessment tools and learning environments for the development of pre-service teachers' professional skills in university-based teacher education (Cochran-Smith & Zeichner, 2005). Such simulations are currently being discussed as promising new measurement systems for improving teachers' professional skills (van Es, Tekkumru-Kisa, & Seago, 2020); however, few such simulations have been created and validated until now. Based on the framework for teaching practice by Grossman et al. (2009), an authentic approximation of practice and a suitable level of cognitive demand can be determined as important dimensions of video-based simulations that allow for practice-oriented teacher learning. This study considered a newly created video-based simulation focusing on teachers' assessment skills regarding students' mathematical argumentation skills as a) an approximation of a real-world professional situation in mathematics classrooms and b) a decomposition of relevant mathematics education content and its adaption to pre-service mathematics teacher assessment skill levels.

Regarding the first research question, the results of this study revealed that the embedded video clips were perceived as authentic representations of practice, matching prior research on videos capturing the authenticity of classroom practice (Brophy, 2004). Despite the use of staged videos, the presented student profiles and their performance on the geometry proof task were perceived as authentic, thus extending prior results on the special medium of staged videos (Böttcher & Thiel, 2018; Piwowar et al., 2017). During the simulation, participants were given the task to assess the students' mathematical argumentation skills by watching teacher–student interactions in multiple video clips and had the possibility to influence the course of working with the simulation by being able to decide on the number of video clips required to finally assess students' argumentation skills. Despite the fact that this environment is quite new for both the participants as well as in (video-based) teacher education research in general, the simulation was perceived as authentic, thus mirroring a real-world professional task. This is also reflected by the external mathematics education experts' qualitative evaluations of the simulation. Therefore, it can be concluded that this newly developed video-based simulation appears to be a useful practice representation of a highly relevant professional task in mathematics teaching. The findings indicated that a valid approximation of practice was created that may be used to help participants acquire practice-oriented knowledge that can effectively be transferred from the simulated environment to real-world practice (Grossman et al., 2009). Furthermore, participants reported being highly cognitively involved in the simulation. In research on simulations, this is considered an important part of participants' presence in a situation, which means that a certain sense of actually being in a virtual environment develops (Schubert et al., 2001). This supports the hypothesis that the participants' behavior in the simulation may mimic real-world situations.

Analyses regarding the cognitive demand of the simulation tasks showed that the basic task of ranking the observed student profiles according to their overall mathematical argumentation skills (*Task 1*) was well-mastered. The results regarding the detailed assessment of the students' argumentation skills (*Task 2*) revealed adequate levels of difficulty in rating the student profiles on the three content dimensions. The targeted medium scoring level was reached, with an adequate distribution around the mean and potential for positive development in the direction of the upper scale of the assessment tasks. As expected, the ranking of the student profiles seemed to be less difficult than providing a detailed assessment of specific aspects of student's argumentation skills. These findings further suggest the internal validity of the simulation, since the implementation of research results from prior mathematics education research regarding commonly observed student skills is in line with our empirical findings of participants' ranking of our simulated students, and an adequate distribution around the mean (Reiss & Ufer, 2009). In addition, these findings indicate that the goal of achieving an adequate level of difficulty is an indicator of whether the decomposition of a complex professional task can be reached (Seidel, Stürmer, Schäfer, & Jahn, 2015; Tekkumru Kisa & Stein, 2015). This indicates that the simulation can be used as an assessment instrument to a) measure pre-service teachers' assessment skills and also b) capture the development of these assessment skills. However, the necessary room for improvement shows that a higher degree of complexity (e.g., realized by inconsistent, simulated student profiles) may indicate an area for future learning once the aspects featured in this simulation have been mastered.

Altogether, the results of this study show that it is possible to develop a video-based simulation that successfully combines two important concepts: i) being perceived as authentic and approximating a real-world classroom situation and ii) cognitive demand not overwhelming pre-service teachers so that participants have the opportunity to acquire practice-oriented knowledge. This also underlines, that both concepts – authenticity and cognitive demand – while related, are not two ends of a continuum, which would limit simulations to either be authentic or have a suitable cognitive demand, but that both concepts differ and can be successfully combined in the creation of video-based simulation. This appears to be a prerequisite for participants to successfully transfer their behavior to the real world.

Although this study's results are promising, some limitations

must be considered. First, the reported study included a sample of 28 participants. This number of participants is relatively small and reduces the value of the study to a certain degree. Therefore, future studies should provide additional evidence by a) collecting qualitative data for further insights into participants' assessment skills; b) gaining insights into learner subgroups and the behavior of the simulation regarding these groups and c) regarding further aspects of validity. Such further aspects of validity can be considered conducting, e.g., comparisons using other instruments and external criteria (e.g. Seidel & Stürmer, 2014; Jahn, 2014). However, the special design of the simulation and its specific, context-related tasks make comparisons difficult. Moreover, the adequacy of the simulation as an assessment and learning instrument for (pre-service) teacher education should be further validated using participants with varying levels of experience to produce more reliable data regarding the simulation's demands at those levels. However, the results regarding the simulation's authenticity can be assumed generalizable to other cohorts, and, as the current difficulty of the detailed analysis was mediocre, there should be some room for both less and more experienced students to use the simulation as well. In particular, additional student profiles—including those showing inconsistent student behavior or other difficulty-generating aspects of real-life students and classroom situations—can be easily added. Finally, the study was mainly based on the framework for teaching practice and considerations regarding the concepts of decomposition and approximation of practice developed by Grossman et al. (2009). However, further research, which might vary conceptual frameworks or concepts of frameworks in systematic ways with the goal of further improving teacher education, is required.

5. Conclusion

The presented study introduced a video-based simulation in the field of teacher education research, focusing on pre-service teachers' assessment skills regarding students' mathematical argumentation skills. By basing the design of the simulation on concepts of the framework for teaching practice (Grossman et al., 2009), a balance between authenticity and cognitive demand emerged as the central design target. In conclusion, this video-based simulation can serve as an evidence-based approximation of practice with the potential to be used in further interventions within teacher education programs. The added-value can be viewed in three ways: First, the study is of conceptual value since it represents an exemplary application of concepts of the framework for teaching practice in the design of a video-based simulation. Balancing authenticity and cognitive demand served as a basis in the development of a simulation in the context of mathematics education. Second, the study is of methodological value since the findings provide further positive evidence regarding the successful implementation of evidence stemming from mathematics education research on observed student argumentation skills in the simulation. Based on the findings of this study, the video-based simulation seems useful for teacher education as part of the valid assessment of practice-oriented knowledge and its acquisition. Third, this simulation might also be used to support teachers' learning when it is enriched with further learning tasks and possibilities for reflection. Therefore, the study has practical implications for teacher education in various educational contexts, since training the assessment of students' mathematical argumentation skills is part of many mathematics teacher education programs worldwide.

Credit author statement

Elias Codreanu: Writing – Original Draft, Methodology, Investigation, Software, Visualization. Daniel Sommerhoff: Writing - Review & Editing, Conceptualization, Methodology. Sina Huber: Formal analysis, Methodology, Software. Stefan Ufer: Supervision, Conceptualization, Methodology. Tina Seidel: Project administration, Funding acquisition, Writing - Review & Editing, Conceptualization.

Acknowledgments

The research was funded by a grant of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) to Tina Seidel, Stefan Ufer, Birgit Neuhaus, and Ralf Schmidmaier (grant number SE 1397/11-1, FOR 2385). The authors thank the participants in our study. Furthermore, the authors thank our research team for their engagement in data collection and analysis.

References

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. Retrieved from https://www.apa.org/ethics/code.

Alles, M., Apel, J., Seidel, T., & Stürmer, K. (2019). How candidate teachers experience coherence in university education and teacher induction: the influence of perceived professional preparation at university and support during teacher induction. *Vocations and Learning, 12*(1), 87–112. https://doi.org/10.1007/s12186-018-9211-5

Alonzo, A. C. (2011). Learning progressions that support formative assessment practices. *Measurement: Interdisciplinary Research & Perspective, 9*(2–3), 124–129. https://doi.org/10.1080/15366367.2011.599629

Bauer, J., & Prenzel, M. (2012). European teacher training reforms. *Science, 336*(6089), 1642–1643. https://doi.org/10.1126/science.1218387

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., … Nickmans, G. (2006). A learning integrated assessment system. *Educational Research Review, 1*(1), 61–67. https://doi.org/10.1016/j.edurev.2006.01.001

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Blomberg, G., Renkl, A., Sherin, M. G., Borko, H., & Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. *Journal of Educational Research Online, 5*(1), 90–114.

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

Böttcher, F., & Thiel, F. (2018). Evaluating research-oriented teaching: A new instrument to assess university students' research competences. *Higher Education, 75*(1), 91–110. https://doi.org/10.1007/s10734-017-0128-y

Brophy, J. E. (2004). *Using video in teacher education. Advances in research on teaching*. Amsterdam: Elsevier JAI.

Chen, C.-H., & Wu, I.-C. (2012). The interplay between cognitive and motivational variables in a supportive online learning system for secondary physical education. *Computers & Education, 58*(1), 542–550. https://doi.org/10.1016/j.compedu.2011.09.012

Chernikova, O., Heitzmann, N., Fink, M., Timothy, V., Seidel, T., & Fischer, F. (2019). Facilitating diagnostic competences in higher education: A meta-analysis in medical and teacher education. *Educational Psychology Review*, 1–40. https://doi.org/10.1007/s10648-019-09492-2

Cochran-Smith, M., & Zeichner, K. M. (2005). Studying teacher education. *The Report of the AERA Panel on Research and Teacher Education*.

Correa, J. M., Martínez-Arbelaiz, A., & Aberasturi-Apraiz, E. (2015). Post-modern reality shock: Beginning teachers as sojourners in communities of practice. *Teaching and Teacher Education, 48*, 66–74. https://doi.org/10.1016/j.tate.2015.02.007

Dankbaar, M. E. W., Alsma, J., Jansen, E. E. H., van Merrienboer, J. J. G., van Saase, J. L. C. M., & Schuit, S. C. E. (2016). An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. *Advances in Health Sciences Education : Theory and Practice, 21*(3), 505–521. https://doi.org/10.1007/s10459-015-9641-x

Darling-Hammond, L., & Bransford, J. (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco: Jossey-Bass.

Derry, S. J., Sherin, M. G., & Sherin, B. L. (2014). Multimedia learning with video. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 785–812). Cambridge: Cambridge University Press.

Dicke, T., Elling, J., Schmeck, A., & Leutner, D. (2015). Reducing reality shock: The effects of classroom management skills training on beginning teachers. *Teaching and Teacher Education, 48*, 1–12. https://doi.org/10.1016/j.tate.2015.01.013

Dieker, L. A., Lane, H. B., Allsopp, D. H., O'Brien, C., Butler, T. W., Kyger, M., et al. (2009). Evaluating video models of evidence-based instructional practices to enhance teacher learning. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children, 32*(2), 180–196. https://doi.org/10.1177/0888406409334202

Pädagogische Psychologie und Entwicklungspsychologie. In Südkamp, A., & Praetorius, A.-K. (Eds.), *Diagnostische kompetenz von lehrkräften: Theoretische und methodische weiterentwicklungen* (Vol. 94), (2017). Münster, New York: Waxmann.

European Commission. (2013). *Supporting teacher competence development: For better learning outcomes*. Retrieved from http://ec.europa.eu/assets/eac/education/experts-groups/2011-2013/teacher/teachercomp_en.pdf.

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M. R., Girwidz, R., Obersteiner, A., … Neuhaus, B. J. (2018). Systematizing professional knowledge of medical doctors and teachers: Development of an interdisciplinary framework in the context of diagnostic competences. *Education Sciences, 8*(4), 207. https://doi.org/10.3390/educsci8040207

Friesen, M., Kuntze, S., & Vogel, M. (2018). Videos, Texte oder Comics? Die Rolle des Vignettenformats bei der Erhebung fachdidaktischer aAnalysekompetenz zum Umgang mit Darstellungen im Mathematikunterricht [Videos, texts or comics? The role of vignette format in the investigation of didactic analysis competences for dealing with presentations in mathematics education]. In J. Rutsch, M. Rehm, M. Vogel, M. Seidenfuß, & T. Dörfler (Eds.), *Effektive Kompetenzdiagnose in der Lehrerbildung* (pp. 153–177). Wiesbaden: Springer Fachmedien.

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., León, V. de, Morrison, D., et al. (2016). Teachers' formative assessment abilities and their relationship to student learning: Findings from a four-year intervention study. *Instructional Science, 44*(3), 267–291. https://doi.org/10.1007/s11251-016-9371-3

Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review, 16*, 41–67. https://doi.org/10.1016/j.edurev.2015.06.001

Gräsel, C., Fischer, F., & Mandl, H. (2000). The use of additional information in problem-oriented learning environments. *Learning Environments Research, 3*(3), 287–305.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055–2100.

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal, 45*(1), 184–205. https://doi.org/10.3102/0002831207312906

Hartmann, S. (1996). The world is a process: Simulations in the natural and social sciences. In R. Hegselmann, U. Mueller, & K. G. Troitzsch (Eds.), *Modelling and simulation in the social sciences from the philosophy of science point of view* (pp. 77–100). Dordrecht: Springer Netherlands.

Heinze, A., & Reiss, K. (2003). *Reasoning and proof: Methodological knowledge as a component of proof competence*. Retrieved from www.lettredelapreuve.itlCERME3PapersiHeinze-paperl.pdf.

Heitzmann, N., Seidel, T., Hetmanek, A., Wecker, C., Fischer, M., & Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research, 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3*(2), 91–98. https://doi.org/10.1016/0742-051X(87)90010-2

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*, 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Jahn, G. K. (2014). *Studien zur Überprüfung der Validität eines Instruments zur Erfassung professioneller Unterrichtswahrnehmung von Lehramtsstudierenden [Studies to test the validity of an instrument for measuring professional vision of preservice teachers]* (Dissertation). Munich: Technical University Munich.

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

Kang, H., & van Es, E. A. (2018). Articulating design principles for productive use of video in preservice education. *Journal of Teacher Education, 5*(1), 1–14. https://doi.org/10.1177/0022487118778549

Karst, K., & Förster, N. (2017). Ansätze der Modellierung diagnostischer Kompetenz - ein Überblick [Approaches to the modeling of diagnostic competences—an overview]. In A. Südkamp, & A.-K. Praetorius (Eds.), *Pädagogische Psychologie und Entwicklungspsychologie: Vol. 94. Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (pp. 19–20). New York: Waxmann: Münster.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38–46. https://doi.org/10.1016/j.tate.2012.10.004

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift Für Psychologie/Journal of Psychology, 216*(2), 61–73. https://doi.org/10.1027/0044-3409.216.2.61

Levin, D. M., Hammer, D., & Coffey, J. E. (2009). Novice teachers' attention to student thinking. *Journal of Teacher Education, 60*(2), 142–154. https://doi.org/10.1177/0022487108330245

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education, 91*, 103059. https://doi.org/10.1016/j.tate.2020.103059

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern [Teachers' diagnostic skills to judge student performance and task difficulty when learning materials include instructional pictures]. *Zeitschrift für Pädagogische Psychologie, 23*(34), 223–235. https://doi.org/10.1024/1010-0652.23.34.223

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. S. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8

Piwowar, V., Barth, V. L., Ophardt, D., & Thiel, F. (2017). Evidence-based scripted videos on handling student misbehavior: The development and evaluation of video cases for teacher education. *Professional Development in Education, 44*(3), 369–384. https://doi.org/10.1080/19415257.2017.1316299

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *The Journal of Educational Research, 106*(1), 64–76.

Praetorius, A.-K., Lipowsky, F., Karst, K., Lazarides, R., & Ittel, A. (2012). Diagnostische Kompetenz von Lehrkräften: Aktueller Forschungsstand, unterrichtspraktische Umsetzbarkeit und Bedeutung für den Unterricht [Diagnostic competences of teachers: Current state of research, teaching practicability and importance for teaching]. In R. Lazarides, & A. Ittel (Eds.), *Differenzierung im mathematisch-naturwissenschaftlichen Unterricht: Implikationen für Theorie und Praxis* (pp. 115–146). Bad Heilbronn: Klinkhardt.

Reiss, K., & Ufer, S. (2009). Was macht mathematisches Arbeiten aus? Empirische Ergebnisse zum Argumentieren, Begründen und Beweisen. *Jahresbericht Der Deutschen Mathematiker-Vereinigung, 2009*(4), 155–177.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84. https://doi.org/10.1002/tea.20163

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology, 6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacogntion, and sense making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York: Simon & Schuster.

Schrader, F.-W. (2017). Diagnostische Kompetenz von Lehrkräften: Anmerkungen zur Weiterentwicklung des Konstrukts [Diagnostic competences of teachers: Comments on the further development of the construct]. In A. Südkamp, & A.-K. Praetorius (Eds.), *Pädagogische Psychologie und Entwicklungspsychologie: Vol. 94. Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (pp. 247–256). New York: Waxmann: Münster.

Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen [Diagnostic competences of teachers: Components and effects]. *Empirische Pädagogik, 1*, 27–52.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence, 10*(3), 266–281.

Seidel, T., Blomberg, G., & Stürmer, K. (2010). Observer: Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht [Observer: Validation of a video-based instrument for investigating the professional perception of lessons]. *Zeitschrift Für Pädagogik, 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

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education, 27*(2), 259–267. https://doi.org/10.1016/j.tate.2010.08.009

Seidel, T., Stürmer, K., Schäfer, S., & Jahn, G. (2015). How preservice teachers perform in teaching events regarding generic teaching and learning components. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(2), 84–96. https://doi.org/10.1026/0049-8637/a000125

Sherin, M. G., Linsenmeier, K. A., & van Es, E. A. (2009). Selecting video clips to promote mathematics teachers' discussion of student thinking. *Journal of Teacher Education, 60*(3), 213–230. https://doi.org/10.1177/0022487109336967

Sommerhoff, D., Ufer, S., & Kollar, I. (2015). Research on mathematical argumentation: A descriptive review of PME proceedings. In K. Beswick, T. Muir, & J. Wells (Eds.), *Psychology of Mathematics Education, 4* pp. 193–200). Hobart, Australia: PME.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teacher judgments on student characteristics and the construct of diagnostic competence]. *Zeitschrift für Pädagogische Psychologie, 19*(1/2), 85–95. https://doi.org/

10.1024/1010-0652.19.12.85
Stokking, K., Leenders, F., Jong, J. de, & van Tartwijk, J. (2003). From student to teacher: Reducing practice shock and early dropout in the teaching profession. *European Journal of Teacher Education, 26*(3), 329–350. https://doi.org/10.1080/0261976032000128175
Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627
Südkamp, A., & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum [Reference-group-effects in a simulated classroom: Direct and indirect judgments]. *Zeitschrift für Pädagogische Psychologie, 23*(34), 161–174. https://doi.org/10.1024/1010-0652.23.34.161
Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz [The simulated classroom: An experimental study on diagnostic competence]. *Zeitschrift für Pädagogische Psychologie, 22*(34), 261–276. https://doi.org/10.1024/1010-0652.22.34.261
Tabak, I., & Kyza, E. A. (2018). Research on scaffolding in the learning sciences: A methodological perspective. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 191–200). New York, London: Routledge Taylor & Francis Group.
Tekkumru Kisa, M., & Stein, M. K. (2015). Learning to see teaching in new ways. *American Educational Research Journal, 52*(1), 105–136. https://doi.org/10.3102/0002831214549452
Tekkumru-Kisa, M., & Stein, M. K. (2017). Designing, facilitating, and scaling-up video-based professional development: Supporting complex forms of teaching in science and mathematics. *International Journal of STEM Education, 4*(1), 27. https://doi.org/10.1186/s40594-017-0087-y
Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., et al. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49*, 36–44.
Ufer, S., Heinze, A., & Reiss, K. (2008). Individual predictors of geometrical proof competence. *PME 32 and PME-NA XXX, 1*(4), 361.
Usiskin, Z. (1982). *Van Hiele levels and achievement in secondary school geometry.* CDASSG Project.
Van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teaching and Teacher Education, 24*(2), 244–276. https://doi.org/10.1016/j.tate.2006.11.005
van Es, Elizabeth A., Tekkumru-Kisa, Miray, & Seago, Nanette (2020). Leveraging the power of video for teacher learning. In Olive Chapman (Ed.), *The international handbook of mathematics teacher education* (pp. 23–54). Leiden; Boston: Brill Sense. https://doi.org/10.1163/9789004418967_002.
Vorderer, P., Wirth, W., Gouveia, F. R., Biocca, F., Saari, T., Jäncke, F., et al. (2004). *MEC spatial presence questionnaire (MEC-SPQ): Short documentation and instructions for application. Report to the European community, Project presence: MEC (IST-2001-37661).* Retrieved from http://www.ijk.hmt-hannover.de/presence.
Weigand, H.-G., Filler, A., Hölzl, R., Kuntze, S., Ludwig, M., Roth, J., et al. (2014). *Didaktik der Geometrie für die Sekundarstufe I (2. Auflage). Mathematik Primarstufe und Sekundarstufe I + II.* Berlin: Springer Spektrum. https://doi.org/10.1007/978-3-642-37968-0
Wigfield, & Eccles. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015
Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*(3), 225–240.