# The Assessment of Teaching Quality Through Classroom Observation – New Approaches for Teacher Education and Research

Dissertation
zur Erlangung des Doktorgrades
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von
Tosca Laetitia Maria Daltoè (geb. Panetta), M.Sc.
aus Horb am Neckar

Tübingen
2024

| | |
|---|---|
| 1. Betreuer: | Prof. Dr. Benjamin Fauth |
| 2. Betreuer: | Prof. Dr. Richard Göllner |
| 3. Betreuer: | Prof. Dr. Ulrich Trautwein |
| | |
| Tag der mündlichen Prüfung: | 14.02.2025 |
| Dekanin: | Prof. Dr. Taiga Brahm |
| Dekan: | Prof. Dr. Dominik Papies |
| | |
| 1. Gutachter: | Prof. Dr. Richard Göllner |
| 2. Gutachter: | Prof. Dr. Andreas Lachner |

# DANKSAGUNG

Diese Dissertation wäre ohne die Unterstützung und Begleitung zahlreicher Personen nicht möglich gewesen und ich möchte mich von ganzem Herzen bei allen bedanken, die dieses Projekt begleitet haben.

An erster Stelle bedanke ich mich bei meinen drei Betreuern Prof. Dr. Richard Göllner, Prof. Dr. Benjamin Fauth und Prof. Dr. Ulrich Trautwein, die mir diese Dissertation ermöglicht haben. Lieber Richard, lieber Ben und lieber Ulrich, danke, dass ihr euch auf das Pilotprojekt „Geteilte Promotionsstelle zwischen Hector-Institut und IBBW“ eingelassen habt und mir so ermöglicht habt, während der gesamten Promotionszeit an zwei Instituten zu arbeiten. Danke für eure enge Betreuung und großartige Förderung. Ich durfte im Rahmen dieser Kooperation professionell und persönlich sehr wachsen. Für die wertvolle Unterstützung bei methodischen Fragen möchte ich dir, Richard, ganz besonders danken.

Außerdem danke ich Prof. Dr. Andreas Lachner herzlich für die Bereitschaft, meine Dissertation als Zweitprüfer zu begutachten und Prof. Dr. Taiga Brahm für die Übernahme des Prüfungsvorsitzes.

Meine tiefe Dankbarkeit gilt außerdem Dr. Evelin Ruth-Herbein und Dr. Ann-Kathrin Jaekel, die mich von Beginn der Promotionszeit an sowohl inhaltlich als auch persönlich begleitet haben. Liebe Evelin, liebe Ann-Kathrin, danke für eure Betreuung, eure Unterstützung und euren Zuspruch, der mich immer wieder zum Weitermachen ermutigt hat!

Ich danke meinen Kolleginnen und Kollegen vom IBBW-Referat 42 und der ganzen Abteilung 4, ganz besonders aber Julia Blank und Jana Caspari für die intensive, schöne und produktive Zusammenarbeit im UFB-Projekt und UFB-Videoprojekt.

Ein besonderer Dank gilt darüber hinaus Linn Hansen, Prof. Dr. Anika Dreher und Prof. Dr. Marita Friesen für die tolle, interdisziplinäre Zusammenarbeit im UFB-Videoprojekt bei der Konzeption und Produktion der geskripteten Unterrichtsvideos im Fach Mathematik. Danke, Linn, für die schönen Drehtage, die wir zusammen an den Schulen verbracht haben. Ein großes Dankeschön geht hierbei auch an all die engagierten Lehrkräften und Schulklassen, die unser Projekt unterstützt haben und sich bereiterklärt haben, mit uns Unterrichtsvideos zu drehen. Die Drehtage wären außerdem nicht möglich gewesen ohne die tatkräftige Unterstützung des Zentrums für Medienkompetenz der Universität Tübingen. Ganz besonders danke ich Kathrin Schumann, Oliver Lichtwald und allen Azubis, die unsere Drehtage begleitet und die Videobearbeitung übernommen haben.

Ich danke Prof. Dr. Peter Gerjets, Dr. Birgit Brucker und Dr. Marc Halfmann für die Ermöglichung der Datenerhebung im Mixed Reality-Lab am IWM Tübingen sowie Isabel Härtl und Cindy Bixel für die tatkräftige Hiwi-Unterstützung. Außerdem danke ich Dr. Tobias Appel und Dr. Philipp Stark für die sehr gute Zusammenarbeit bei der Eye-Tracking-Studie.

Bei Prof. Dr. Kirsti Klette von der Universität Oslo und der Arbeitsgruppe rund um das QUINT-Netzwerk möchte ich mich für den schönen Forschungsaufenthalt, den ich Anfang 2024 in Oslo verbringen durfte, bedanken. Ich konnte während meines Aufenthalts viel über videobasierte Unterrichtsqualitätsforschung und Beobachtungsratings lernen.

Außerdem danke ich der Tübinger Graduiertenschule LEAD für die zahlreichen Retreats, das tolle Forschungsnetzwerk und die finanzielle Förderung von Konferenzreisen, meines Forschungsaufenthaltes sowie Proofreadings von Manuskripten.

Neben der konkreten inhaltlichen Zusammenarbeit wurde ich von vielen weiteren Menschen auf meinem Promotionsweg begleitet und unterstützt. Ich danke meinen „Roomies“ Wy Ming, Fitore, Alex, Mapi und Lisa-Marie für die schöne und lustige Bürozeit und für den Zuspruch, gerade auf den letzten Metern der Promotionszeit. Außerdem danke ich lieben Kolleginnen, die zu Freundinnen wurden und lieben Freundinnen, die auch zu Kolleginnen wurden und so allesamt meinen Promotionsweg besonders eng begleitet haben: Danke Aki, Babette, Fitore und Hannah sowie Franzi, Totti, Emely und Ronja. Danke, dass ihr mich durch alle Höhen und Tiefen begleitet und getragen habt!

Zu guter Letzt danke ich den Menschen, die mich schon am längstem im Leben begleiten: Danke an meine Familie und meine Freundinnen und Freunde. Ihr steht mir stets mit Rat und Tat zur Seite und sorgt im richtigen Moment für die notwendige Ablenkung. Danke, dass es euch gibt! Und weil keine Worte für deine Liebe und Unterstützung ausreichen: einfach nur danke, Robin!

# ABSTRACT

The quality of classroom teaching plays a central role in students' academic achievement and motivation (Burroughs et al., 2019). A model frequently used in German-speaking countries defines three basic dimensions of teaching quality: cognitive activation, student support, and classroom management (Klieme et al., 2009). To promote teaching quality based on these dimensions, reliable and valid methods for assessing teaching quality are required. One central method is classroom observation by external observers (O'Leary, 2020). Classroom observation, often conducted using classroom videos, allows for systematic and detailed analysis of teaching in both research and practice contexts. Although the perspective of external observers offers numerous advantages, there are clear limitations in capturing teaching quality through observer ratings, such as the difficulty of observing certain teaching-quality aspects from an outside perspective or the influence of idiosyncratic rater characteristics on the ratings (Fauth et al., 2020; Göllner et al., 2016). Previous research has shown that these challenges are often reflected in the limited psychometric quality of observer ratings of teaching quality (e.g., Kelly et al., 2020). Given the potential of (video-based) classroom observations for teacher education and educational research, it is important to investigate under which conditions external observers can provide the most accurate assessments of teaching quality possible.

The aim of the present dissertation was to examine conditions for accurate observation-based assessments of teaching quality. To this end, a theoretical model of observers' assessment accuracy in classroom observations was developed, based on Funder's (1995) model of judgment accuracy in personality assessment. The proposed model outlines a four-step process necessary for accurate assessments of teaching quality: relevance, availability, detection, and utilization of behavioral cues in the teaching. Additionally, the model identifies two key factors that can influence this assessment process: the observers' disposition and the observation environment in which the classroom observation takes place. This dissertation includes three empirical studies, each representing a different approach to examining observer ratings of teaching quality and systematically investigating aspects of the proposed theoretical model of observers' assessment accuracy in classroom observations.

The first study (*Can teachers be trained to provide valid teaching-quality ratings?*) focused on the role of the observers' disposition in providing observer ratings of teaching quality, specifically the level of training in the use of standardized classroom observation systems. In this study, a rater training for in-service teachers in the use of a standardized

classroom observation system was followed across five time points. The study examined how the reliability of teaching-quality ratings provided by teachers, based on classroom videos, developed throughout the training. Following the training, a validation study was conducted to assess the validity of the ratings provided by the trained teachers. The results showed that the teachers were able to rate the observed teaching-quality aspects with increasing agreement over the course of the training, with varying developments of interrater agreement depending on the aspect being assessed. Furthermore, three validity arguments for using the trained teachers' ratings were derived: increased agreement with expert ratings, a fitting factor structure, and high convergent correlations with ratings from a comparable classroom observation system. With these findings, the first study emphasizes the importance of rater training in using standardized classroom observation systems for accurate assessments of teaching quality from the perspective of external observers.

The second study (*Immersive insights: Unveiling the impact of 360-degree videos on pre-service teachers' classroom observation experiences and teaching-quality ratings*) investigated the impact of the observation environment, comparing two different video environments: a traditional video environment on a computer and an immersive 360-degree video environment using virtual reality (VR) headsets. This study examined the classroom observation experience and resulting observer ratings of teaching quality of pre-service teachers in both video environments. The results indicated that in the immersive 360-degree environment, the pre-service teachers felt more cognitively, affectively, and physiologically involved in the classroom teaching and reported higher motivation, without reporting a higher mental workload. Teaching-quality ratings in both video environments were largely similar, except for one aspect of cognitive activation (*focus on key concepts*), which was rated more accurately in the immersive 360-degree video environment. This study highlights the potential of immersive 360-degree classroom videos for teacher education and the assessment of teaching quality.

The third study (*Connecting gaze behavior and ratings of teaching quality*) was based on the assumption of the proposed theoretical model that the observer must successfully detect the quality-relevant classroom events for an accurate assessment of teaching quality. Using eye-tracking data from classroom observations, this study examined the relationship between the gaze behavior of pre-service teachers and their observer ratings of teaching quality. It also compared these relationships between classroom observations in traditional and immersive 360-degree video environments. The results indicated that visual attention to specific areas in

the classroom, as well as pupil diameter during quality-relevant classroom events, were related to the accuracy of observer ratings of teaching quality, with these relationships being particularly pronounced in immersive 360-degree video environments. Thus, the third study highlights the potential of using eye-tracking data as process data from classroom observations to deepen our understanding of the perceptual processes underlying observer ratings of teaching quality in classroom observations.

The present dissertation provides novel insights into the conditions for successfully assessing teaching quality through classroom observations, thereby enhancing our understanding of observer ratings of teaching quality. By proposing a theoretical model of observers' assessment accuracy in classroom observations and presenting innovative research approaches—such as the systematic investigation of rater training, the use of immersive 360-degree classroom videos, and the integration of eye-tracking data as process data from classroom observations—this dissertation makes significant theoretical, methodological, and empirical contributions for applying classroom observations in teacher education, professional development, and classroom research.

# ZUSAMMENFASSUNG

Unterrichtsqualität spielt eine zentrale Rolle für die Leistungs- und Motivationsentwicklung von Schülerinnen und Schülern (Burroughs et al., 2019). Ein insbesondere im deutschsprachigen Raum häufig genutztes Modell definiert drei Basisdimensionen von Unterrichtsqualität: Kognitive Aktivierung, Konstruktive Unterstützung und Strukturierte Klassenführung (Klieme et al., 2009). Um eine gezielte Förderung der Unterrichtsqualität anhand dieser Dimensionen zu ermöglichen, bedarf es zuverlässiger und valider Methoden zur Erfassung von Unterrichtsqualität. Eine zentrale Methode ist die Unterrichtsbeobachtung durch externe Beobachtende (O'Leary, 2020). Die Unterrichtsbeobachtung, die häufig anhand von Unterrichtsvideos durchgeführt wird, ermöglicht eine systematische und detaillierte Analyse des Unterrichts in Forschungs- und Praxiskontexten. Obwohl die Perspektive externer Beobachtender zahlreiche Vorteile bietet, zeigen sich auch klare Grenzen bei der Erfassung von Unterrichtsqualität durch Beobachtungsratings, wie etwa dass bestimmte Merkmale von Unterrichtsqualität schwer von außen beobachtbar sind oder dass Beobachtungsratings durch idiosynkratische Ratermerkmale beeinflusst werden (Fauth et al., 2020; Göllner et al., 2016). Bisherige Forschung hat gezeigt, dass sich diese Herausforderungen auch in oft eingeschränkter psychometrischer Qualität von Unterrichtsbeobachterratings niederschlagen (z. B. Kelly et al., 2020). Aufgrund des großen Potenzials (videobasierter) Unterrichtsbeobachtungen für Lehrkräftebildung und Unterrichtsforschung sollte daher verstärkt untersucht werden, unter welchen Bedingungen die Perspektive externer Beobachtender möglichst akkurate Urteile über Unterrichtsqualität erlaubt.

Das Ziel der vorliegenden Dissertation war es, Bedingungen für eine erfolgreiche, beobachtungsbasierte Erfassung von Unterrichtsqualität zu untersuchen. Hierfür wurde ein theoretisches Modell zur Urteilsgenauigkeit externer Beobachtender in Unterrichtsbeobachtungen entwickelt, das auf einem Modell der Urteilsgenauigkeit von Funder (1995) aus dem Bereich der Persönlichkeitseinschätzung basiert. Das vorgeschlagene Modell beschreibt vier Prozessschritte, die für eine akkurate Einschätzung von Unterrichtsqualität durch Beobachtungsratings notwendig sind: Relevanz, Verfügbarkeit, Erkennung sowie Nutzung behavioraler Hinweise aus dem Unterrichtsgeschehen. Zusätzlich definiert das Modell zwei zentrale Aspekte, die den Einschätzungsprozess von Unterrichtsqualität beeinflussen können: die individuelle Disposition der Beobachter und das Setting, in dem die Unterrichtsbeobachtung stattfindet. Die vorliegende Dissertation beinhaltet drei empirische Studien, die unterschiedliche Ansätze zur Untersuchung von Beobachtungsratings repräsentieren und dabei

systematisch Teilaspekte des vorgeschlagenen theoretischen Modells der Urteilsgenauigkeit externer Beobachtender in Unterrichtsbeobachtungen adressieren.

Die erste Studie (*Can teachers be trained to provide valid teaching-quality ratings?*) nahm die Rolle der individuellen Disposition von Beobachtern zur Bereitstellung von Unterrichtsbeobachtungsratings in den Blick, wobei konkret das Level an Training in der Nutzung von standardisierten Unterrichtsbeobachtungsinstrumenten im Fokus stand. In dieser Studie wurde ein Rater-Training für Lehrkräfte zur Nutzung eines standardisierten Unterrichtsbeobachtungsinstruments über fünf Messzeitpunkte hinweg begleitet. Es wurde untersucht, wie sich die Reliabilität von Unterrichtsbeobachtungsratings im Verlauf des Trainings entwickelte. Nach dem Training schloss sich eine Validierungsstudie an, in der die Validität der Ratings der trainierten Lehrkräfte in den Blick genommen wurde. Die Ergebnisse zeigten, dass die Lehrkräfte die erfassten Unterrichtsqualitätsmerkmale im Verlauf der Schulung mit zunehmender Übereinstimmung einschätzen konnten, wobei sich je nach erfasstem Aspekt unterschiedliche Entwicklungsverläufe in der Beobachtungs-übereinstimmung zeigten. Außerdem konnten drei Validitätsargumente für die Nutzung der Ratings geschulter Lehrkräfte abgeleitet werden: eine erhöhte Übereinstimmung mit Expertenratings, eine passende Faktorstruktur und hohe konvergente Zusammenhänge mit Ratings eines vergleichbaren Beobachtungsinstruments. Die erste Studie dieser Dissertation betont die Bedeutsamkeit von Rater-Trainings in der Nutzung standardisierter Unterrichtsbeobachtungsinstrumente für eine akkurate Erfassung von Unterrichtsqualität aus der Perspektive externer Beobachtender.

Die zweite Studie (*Immersive insights: Unveiling the impact of 360-degree videos on pre-service teachers' classroom observation experiences and teaching-quality ratings*) untersuchte den Einfluss des Settings der Unterrichtsbeobachtung, wobei konkret zwei verschiedene Unterrichtsvideoumgebungen miteinander verglichen wurden: eine traditionelle Videoumgebung am PC und eine immersive 360-Grad-Videoumgebung mit Virtual Reality (VR)-Brillen. Die Studie untersuchte das Unterrichtserleben und resultierende Unterrichtsbeobachtungsratings von Lehramtsstudierenden in beiden Videoumgebungen. Die Ergebnisse zeigten, dass sich die Lehramtsstudierenden in der immersiven 360-Grad-Umgebung kognitiv, affektiv und physiologisch stärker in den Unterricht eingebunden fühlten und eine höhere Motivation aufwiesen, ohne einen höheren mentalen Workload zu berichten. Die Einschätzungen der Unterrichtsqualität fielen in beiden Videoumgebungen weitgehend gleich auf, bis auf einen Aspekt der kognitiven Aktivierung (*Verständnisorientierung*), der in

der immersiven 360-Grad-Videoumgebung akkurater eingeschätzt wurde. Diese Studie hebt das Potenzial immersiver 360-Grad-Unterrichtsvideos für die Lehrkräftebildung und die Erfassung von Unterrichtsqualität hervor.

Die dritte Studie (*Connecting gaze behavior and ratings of teaching quality*) basierte auf der Grundannahme des vorgeschlagenen theoretischen Modells, dass qualitätsrelevante Unterrichtsereignisse für eine akkurate Einschätzung der Unterrichtsqualität zunächst vom Beobachter erfolgreich erkannt werden müssen. Anhand von Blickbewegungsdaten aus Unterrichtsbeobachtungen untersuchte diese Studie, wie das Blickverhalten von Lehramtsstudierenden mit deren Unterrichtsbeobachtungsratings zusammenhing. Zudem wurden diese Zusammenhänge zwischen Unterrichtsbeobachtungen in einer traditionellen und einer immersiven 360-Grad-Videoumgebung verglichen. Die Ergebnisse wiesen darauf hin, dass die visuelle Aufmerksamkeit auf bestimmte Bereiche im Unterricht sowie der Pupillendurchmesser während qualitätsrelevanter Unterrichtsereignisse mit der Genauigkeit von Unterrichtsbeobachtungsratings zusammenhängen, wobei sich diese Zusammenhänge vor allem in immersiven Videoumgebungen zeigten. Damit weist die dritte Studie dieser Dissertation auf das Potenzial der Nutzung von Blickbewegungsdaten als Prozessdaten aus Unterrichtsbeobachtungen hin, um Wahrnehmungsprozesse, die der beobachtungsbasierten Einschätzung von Unterrichtsqualität zugrunde liegen, tiefergreifend zu verstehen.

Die vorliegende Dissertation liefert neue Erkenntnisse über die Bedingungen für eine erfolgreiche Erfassung der Unterrichtsqualität durch Unterrichtsbeobachtungen und trägt damit zu einem vertieften Verständnis von Unterrichtsbeobachtungsratings bei. Durch die Entwicklung eines theoretischen Modells zur Urteilsgenauigkeit externer Beobachtender in Unterrichtsbeobachtungen und die Präsentation innovativer Forschungsansätze, wie der systematischen Untersuchung eines Rater-Trainings, der Verwendung immersiver 360-Grad-Unterrichtsvideos und der Integration von Blickbewegungsdaten als Prozessdaten aus Unterrichtsbeobachtungen, leistet diese Dissertation bedeutende theoretische, methodische und empirische Beiträge zur Anwendung von Unterrichtsbeobachtungen in der Lehrkräftebildung und der Unterrichtsforschung.

# CONTENTS

# 1

# INTRODUCTION AND THEORETICAL BACKGROUND

# 1 Introduction and Theoretical Background

"Do you remember Mrs. Cortés' Spanish class? Her teaching was excellent, and I learned so much." Conversations like this often pop up among school friends. We all carry memories of classes where the teaching stood out—sometimes for better, sometimes for worse. But what defines good or bad teaching? And what specific characteristics of teaching influence our assessments of teaching quality? These questions are not only pertinent to our personal reflections on school teaching but also crucial for researchers and practitioners who discuss, assess or evaluate the quality of classroom teaching.

Teaching is a highly complex interaction between teachers and students (Praetorius & Charalambous, 2023). To better understand the complexity of teaching, educational research has focused for many years on what characterizes effective teaching and how teaching quality can be assessed. Over time, several teaching characteristics have been identified that are predictive for students' learning outcomes and motivation (Hattie, 2009; Klieme et al., 2009; Seidel & Shavelson, 2007). In addition, methods have been developed to reliably and validly assess the quality of teaching. One of the key approaches for assessing teaching quality is classroom observation by external observers (Clausen, 2002; Fauth et al., 2020; Göllner et al., 2016; O'Leary, 2020).

There is a large body of literature on assessing teaching quality through classroom observations (see e.g., overviews by Kelly et al., 2020; Klette & Blikstad-Balas, 2018; Martinez et al., 2016). The observer perspective on teaching quality has been used in both practical contexts and educational research. In practice, classroom observations are not only used for teaching evaluations but also in teacher education and professional development, where classroom observation often takes place using classroom videos (Gaudin & Chaliès, 2015; Martinez et al., 2016; Taut & Rakoczy, 2016). In this context, for example, teacher trainings including video observations proved to be an effective method to foster professional competencies of (pre-service) teachers, such as the professional vision of classroom management (Weber et al., 2018) or feedback competence (Prilop et al., 2020). In educational research, observation-based assessments of teaching quality in video studies, such as the Third International Mathematics and Science Study (TIMSS; Hiebert & Stigler, 2000), have contributed significantly to our understanding of teaching quality, offering insights into its conceptualization (Klieme et al., 2009; Pianta & Hamre, 2009) or revealing strengths and weaknesses regarding various dimensions of teaching quality across countries (White & Klette, 2023).

The observer perspective on classroom teaching holds great potential in both research and practice settings, though it also presents certain challenges. For example, trained observers can serve as independent experts, promising a more objective view of teaching quality (Petko et al., 2003; Storms, 1973), but at the same time not all behaviors crucial to teaching quality are directly observable (Fauth et al., 2020) and observers may be subject to bias, especially in brief classroom video sequences (Praetorius, 2014). These challenges of the observer perspective on teaching quality also become evident in research findings, where studies showed limited psychometric quality of observer ratings (e.g., Kelly et al., 2020; Praetorius et al., 2012; White & Klette, 2024).

The consistent result pattern around the limited psychometric quality of observer ratings raises an important yet open question for using classroom observations in research and practice: Under what conditions can observers provide accurate assessments of teaching quality in (video-based) classroom observations? The present dissertation aims to tackle this question from three different perspectives. Based on a proposed model of observers' assessment accuracy in classroom observations (adapted from a model of judgement accuracy in personality judgement; Funder, 1995), this dissertation addresses three different approaches to gain a deeper understanding of observer ratings of teaching quality.

The first approach to understanding observer ratings of teaching quality more profoundly focuses on the observers themselves and their individual disposition to provide reliable and valid ratings of teaching quality (Bell et al., 2014; White & Ronfeldt, 2024). To do so, Study 1 of this dissertation systematically investigates observer ratings of teaching quality throughout a rater training for in-service teachers in using a standardized classroom observation system created for feedback on teaching quality in school practice (Fauth et al., 2021).

The second approach to understanding observer ratings of teaching quality more profoundly focuses on the effects of the video environment. Classroom observations always take place in a specific environment, which can be the actual classroom or different forms of video-based environments. In recent years, technological innovations have increased the possibility of recording and displaying videos (e.g., Snelson & Hsu, 2020), which also impacts the possibilities for video-based classroom observation. As observer ratings of teaching quality have been shown to differ depending on the observation environment (Curby et al., 2016; Jentsch et al., 2024; Paulicke et al., 2019; Wyss et al., 2023), Study 2 of this dissertation compares observers' classroom observation experiences and observer ratings of teaching quality between two video-based observation environments: a traditional classroom video

environment on a computer and an immersive 360-degree classroom video environment using head-mounted displays (HMD), so-called virtual reality (VR) glasses.

The third approach to understanding observer ratings of teaching quality more profoundly is to focus on the observers' perceptual processes during classroom observation via eye-tracking technology. This approach is based on the assumption that observers need to detect quality-relevant events in the classroom to interpret them correctly and draw the correct inferences about the quality of specific aspects of teaching quality (König et al., 2022; Santagata et al., 2021). For this reason, besides investigating only the ratings resulting from classroom observations, taking a closer look at the process of observation itself may offer valuable insights into how observers look at the teaching in their assessment process. For this reason, Study 3 of this dissertation investigates how observers' gaze behavior is connected to the accuracy of their teaching-quality ratings, again in the context of different classroom video environments.

The present dissertation is structured as follows: Chapter 1 presents the theoretical background underlying this dissertation. After highlighting the importance of teaching for learning and conceptualizing teaching quality as the construct of interest and addressing its assessment (Chapter 1.1), I take a closer look at the observer perspective on teaching quality (Chapter 1.2). Here, I summarize promises but also conceptual and empirical limitations of observer ratings and highlight the importance of rater trainings as a quality procedure to build the observers' disposition to provide accurate ratings of teaching quality. In this chapter, I start to stepwise propose a model of observers' assessment accuracy in classroom observations, as the theoretical foundation to locate my dissertation studies. Following this, I present the specific characteristics of classroom observation in video-based environments (Chapter 1.3), where I also introduce 360-degree videos as an innovative approach to create immersive video environments for classroom observations as well as the power of eye-tracking for assessing process data of classroom observations. To conclude the introduction and theoretical background, I summarize the proposed theoretical model of observers' assessment accuracy in classroom observations along with all its derived components (Chapter 1.4). Based on the theoretical background, Chapter 2 describes this dissertation's aims and research questions. Chapters 2 to 5 then present the empirical studies addressing my research questions. In Chapter 6, I close with a general discussion. Discussing the findings of my three dissertation studies, I highlight the contributions of the respective study and reflect on the significance of each of the three approaches in the light of the proposed theoretical model of observers' assessment accuracy in classroom observations (Chapter 6.1). Subsequently, I discuss the strengths and

limitations of this work (Chapter 6.2), as well as implications for research and practice (Chapter 6.3) before closing with a general conclusion (Chapter 6.4).

## 1.1 Teaching Quality and Why it Matters

"Differentiated Instruction Made Practical", "Teaching Students to Ask Their Own Questions: Best Practices in the Question Formulation Technique", "Teaching 4.0—Challenging Student Behavior—From Intervention to Prevention". These are exemplary titles of professional development programs for teachers currently advertised online (Harvard Graduate School of Education, 2024; Lehrkräftefortbildung Baden-Württemberg, 2024). It is no coincidence that teaching is a central component in teacher education and professional development programs, as it is clear: Teaching matters! Teaching is the "core business" of school (Helmke, 2009; Reusser, 2008) and its quality affects students' educational success (e.g., Blömeke et al., 2022). Due to its importance, teaching quality, also referred to as instructional quality, has been a construct of interest in educational research for several decades (Hattie, 2009; Klieme, 2019; Wang et al., 1993). In the following, I start with describing the role of teaching for students' learning outcomes more closely (Chapter 1.1.1), before summarizing current approaches to conceptualize and assess teaching quality (Chapter 1.1.2).

### 1.1.1 The Role of Teaching for Learning

"Does teaching make a difference?" (Brophy & Good, 1986) was a dominant question in educational research for a long time. The significance of teaching, which can be defined as the teacher-initiated process of fostering classroom interactions to create learning opportunities for students (Hiebert & Stigler, 2023), has not always been fully recognized. Whereas today, most researchers and practitioners would clearly agree with the importance of teaching for learning, the answer to this question has not always been clear.

Some of the early empirical studies focusing on determinants of students' learning concluded that academic achievement is primarily influenced by school-independent factors, such as students' individual social background or their cognitive predispositions, with school-related factors having little impact (Coleman et al. 1966; Jencks et al., 1972). However, with improved methodological approaches to accounting for instructional factors, studies began to provide increasing evidence that teaching does play a significant role in students' learning (Brophy & Good, 1986; Scheerens & Bosker, 1997; Wang et al., 1993). The first theoretical model including instructional aspects determining academic achievement was Carroll's model of school learning (Carroll, 1963). Carroll (1963) assumed that academic achievement is based on the ratio of available learning time to the time needed to learn. As relevant factors, he included the opportunity to learn (time available to learn), the ability to understand instruction,

the quality of instructional events and the learners' perseverance (time the learner is willing to spend learning). Later, Bloom (1976) and Walberg (1981) also included the quality of teaching in their theoretical models to explain students' academic achievement. During this time, classroom research was dominated by the process-product paradigm, shaped by behaviorism (Gage & Needels, 1989; Gräsel & Göbel, 2011). Within the process-product paradigm, research focused on investigating which aspects of teaching (*process*) are associated with students' educational outcomes (*product*). However, studies within the process-product paradigm were often merely correlational and primarily neglected contextual factors of learning (Gräsel & Göbel, 2011). A popular advancement of the process-product paradigm, particularly in the German-speaking context, is the *Utilization of Learning Opportunities Model* (Angebots-Nutzungs-Modell; Vieluf & Klieme, 2023). This model assumes that teaching represents an offer made by the teacher that does not directly influence students' learning. Instead, it must be utilized in light of various contextual variables, such as social background or classroom composition. Having these advanced assumptions about non-directional effects on students' learning, research designs improved and went beyond the mostly correlative associations of the process-product paradigm, accounting for contextual factors of teaching (Scheerens et al., 2007; Seidel & Shavelson, 2007). Examples of influential research studies on the determinants of successful learning are the meta-analysis by Seidel and Shavelson (2007) and the synthesis of over 800 meta-analyses by John Hattie (2009). Seidel and Shavelson (2007) found systematic effects of teaching on learning and made clear that the research design matters when investigating the role of teaching for learning. For example, effects of teaching on learning become particularly evident in experimental research designs (Seidel & Shavelson, 2007). Hattie (2009) found that especially teaching-related aspects predict students' learning outcomes and thus concluded that teaching makes a difference for learning (Hattie, 2009). Although Hattie's methodology of synthesizing effect sizes from a wide range of studies has been heavily criticized (e.g., Rømer, 2019), his work remains important for understanding the determinants of students' learning. In recent years, Burroughs et al. (2019) reviewed the literature on teaching effectiveness and student outcomes. The authors summarize that students' achievement is related to several teaching-related factors, such as teachers' support, teaching preparation, or provision of opportunity to learn (Burroughs et al., 2019).

Summing up, research has indicated that we can confidently contradict Coleman et al. (1966) and Jencks et al. (1972) and state: High-quality teaching does significantly impact students' learning (Blömeke et al., 2022; Blömeke & Olsen, 2019).

### 1.1.2 Conceptualization and Assessment of Teaching Quality

As the importance of teaching for learning became evident, research explored ways to conceptualize teaching quality and developed methods for its assessment (Panayiotou et al., 2021; Senden et al., 2022). Berliner (2005) highlights that a clear-cut definition of teaching quality is impossible, as cultural and societal perspectives inevitably shape any assessment of teaching quality. He distinguishes two fundamentally different components of high-quality teaching: *Good Teaching*, which refers to the normative aspect of teaching shaped by culturally and socially dominant expectations and values, and *Effective Teaching*, which encompasses the objective, descriptive aspect focused on achieving specific educational goals, such as acquiring knowledge or skills (Berliner, 2005). Teaching quality can be described as a combination of these two components. In empirical research on teaching quality, instructional practices are systematically evaluated for their impact on students, implying a view of teaching quality based on Berliner's concept of effective teaching (Klieme, 2019). In this context, teaching quality is defined as empirically observable features of classroom interactions associated with students' development in terms of achieving educational goals (Klieme, 2019). However, Sauerwein and Klieme (2016) note that prevailing understandings of teaching quality are always grounded in normative assumptions. Thus, both components of high-quality teaching, as outlined by Berliner (2005), influence conceptualizations of teaching quality in educational research.

Conceptualizing teaching quality, research for a long time focused on identifying aspects of teaching that constitute effective teaching in terms of predicting students' outcomes, such as academic achievement or motivation. Several researchers have published lists of effective teaching characteristics, which have substantially impacted both research and educational practice (e.g., Brophy, 2000; Danielson, 2007; Helmke, 2004; Meyer, 2003). Groundbreaking work on conceptualizing teaching quality in the German-speaking context has emerged from the research group led by Eckhard Klieme (Klieme et al., 2001). Within the context of the TIMSS video study (Baumert et al., 1997), Klieme and colleagues conducted exploratory factor analyses to identify a structure to describe the quality of teaching. Here, three basic dimensions for a generic description of teaching quality emerged: *cognitive activation*, *student support*, and *classroom management*. The three dimensions identified by Klieme et al. (2001) provided the basis for the "Model of the Three Basic Dimensions of Teaching Quality", which is currently the widely established conceptualization for teaching quality, especially in German-speaking contexts (Klieme et al., 2009; Praetorius et al., 2018). Internationally, however, there are conceptualizations of teaching quality incorporating three similar

dimensions. For example, the teaching-quality model by Pianta and Hamre (2009) differentiates between the three teaching quality dimensions *instructional support*, *emotional support*, and *classroom organization*.

The Model of the Three Basic Dimensions of Teaching Quality (TBD) is widely used to describe teaching quality in both research and practice. It serves as the central theoretical framework for teaching quality in the present dissertation. For this reason, I describe the theoretical assumptions behind the TBD in more detail in the following.

The basic dimension of *cognitive activation* is grounded in a cognitive-constructivist theory of learning (Praetorius et al., 2018). Cognitive activation refers to stimulating learners to actively engage with content to achieve a deep understanding of this content (Kunter & Trautwein, 2013). This basic dimension has been especially researched in mathematics teaching, where several teaching characteristics have been identified as aspects of cognitively activating teaching. For instance, profound cognitive activation of learners occurs through exploration and connections to their prior knowledge (Baumert & Köller, 2000) or by presenting highly challenging problems and questions. Cognitively activating teaching requires high-level cognitive processing and metacognition, which are assumed to encourage an elaborate engagement with the learning material (Baumert et al., 2010; Lipowsky et al., 2009). Consequently, the hypothesized underlying mechanism of cognitive activation is the depth of processing, through which cognitively activating teaching is assumed to positively impact students' achievement (Klieme et al., 2009; Praetorius et al., 2020).

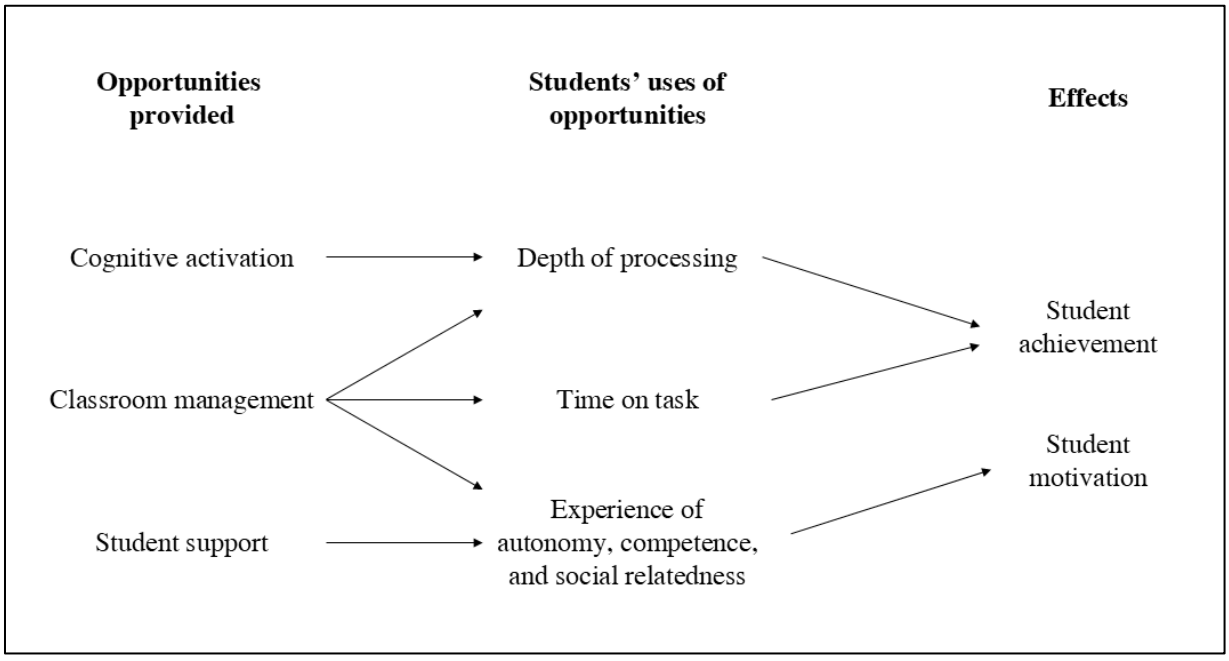
The basic dimension of *student support* is rooted in research on school and classroom climate (Clausen, 2002) and draws on theoretical principles of the *Self-Determination Theory of Motivation* (Deci & Ryan, 2000). Student support refers to how teachers assist students when they encounter difficulties understanding content and the degree to which teacher-student interactions are characterized by respect and appreciation (Kunter & Trautwein, 2013). Student support can be divided into two subfacets: a cognitive facet and an emotional-motivational facet (Kleickmann et al., 2020). The cognitive facet of student support includes, for example, constructive feedback (Hattie & Timperley, 2007; Wisniewski et al., 2020) or individualized support adapted to students' individual learning levels (Dumont, 2019). The emotional-motivational facet, by contrast, involves a positive teacher-student relationship characterized by mutual respect and recognition (Kunter & Trautwein, 2013). Therefore, the hypothesized mechanism behind the basic dimension of student support is the experience of autonomy,

competence, and social relatedness, which is assumed to enhance students' motivation (Klieme et al., 2009; Praetorius et al., 2020).

The basic dimension of *classroom management* builds on Carroll's (1963) time-on-task hypothesis. It encompasses the actions and strategies teachers use to coordinate the complex dynamics of the classroom, intending to minimize disruptions and optimize learning time (Kunter & Trautwein, 2013; Kunter & Voss, 2013). Effective classroom management aligns available teaching time with time spent engaging with learning content. The hypothesized mechanism underlying this dimension is the efficient use of time on task, which is expected to foster deeper cognitive processing and a stronger sense of competence. Consequently, effective classroom management is assumed to enhance both students' achievement and motivation (Klieme et al., 2009; Praetorius et al., 2020). Figure 1 provides a graphical representation of the assumed relationship between the TBD and student outcomes.

**Figure 1**

*Assumed Relations Between the Three Basic Dimensions of Teaching Quality and Student Outcomes (Praetorius et al., 2020, p. 20, adapted from Klieme et al., 2009)*




In addition to the conceptualization of teaching quality, determining how to effectively assess it remains a primary focus of educational research (Fauth et al., 2020; Göllner et al., 2016; Senden et al., 2022). Accurate methods for assessing teaching quality are crucial, as teaching assessments inform policy, practice, and research (White & Klette, 2023). In policy and practice, for instance, teaching evaluations can influence a teacher's career path, for

example by rewarding teachers for good teaching evaluations (e.g., Boyd et al., 2008; Rodriguez et al., 2020). In research on teaching and learning, the way to assess teaching quality in a study shapes the findings, thereby affecting how this construct is understood. Such findings can, in turn, inform the development of teacher education curricula (e.g., Hollins, 2011) and the design of measurement instruments for future studies (e.g., Klieme et al., 2009; Pianta & Hamre, 2009). For this reason, reliable and valid assessments of teaching quality are essential prerequisites for effective teaching-quality development, ultimately impacting students' learning outcomes.

Typically, three central perspectives are used to assess teaching quality: teachers' self-assessment, student ratings, and classroom observations by external observers (Clausen, 2002; Fauth et al., 2020; Göllner et al., 2016). In recent years, the three well-established perspectives have been joined by automated approaches to assess teaching quality (Foster et al., 2024), such as automated assessment of hand-raising behavior as an indicator of students' engagement (Bühler et al., 2023) or the automated assessment of encouragement and warmth as an indicator of a supportive classroom climate (Hou et al., 2024). Empirical studies on the interrelationship of the different perspectives to assess teaching quality tend to show low levels of agreement, depending on the teaching-quality dimensions assessed. Whereas aspects related to classroom management overall are assessed with higher agreement between perspectives, aspects of cognitive activation or student support tend to be rated very differently across different perspectives (Clausen, 2002; Fauth et al., 2014; Kunter & Baumert, 2007; Wagner et al., 2016). These perspective-specific deviations in teaching-quality ratings suggest that not all aspects of teaching quality can be adequately assessed from each of the three perspectives.

Due to the individual background and varying involvement in the teaching process, each perspective to assess teaching quality has its own promises and pitfalls when evaluating specific aspects of teaching quality (for an overview, see Fauth et al., 2020). However, the outsider perspective of external observers as experts of teaching quality has had a special role in research on teaching quality. Observer ratings have even been considered the gold standard for assessing teaching quality (Helmke, 2009). Despite the considerable potential that an independent observer perspective brings to the assessment of teaching quality, empirical findings have repeatedly pointed to the limitations of assessing teaching quality through observer ratings (Bell et al., 2014; Kelly et al., 2020; Praetorius et al., 2012; White & Klette, 2024). Thus, it becomes clear that observer ratings offer both promises and pitfalls in assessing teaching quality. As the

observer perspective on teaching quality is the central objective of this dissertation, it will be examined in detail in the following chapter.

## 1.2 The Observer Perspective on Teaching Quality

"Please observe the teaching sequence in the following video carefully and assess the quality of teaching using the rating items of the Classroom Assessment Scoring System", could be a prototypical instruction for a rater using the Classroom Assessment Scoring System (CLASS), a widely used observation system to assess teaching quality in classroom research and practice contexts (Pianta et al., 2008). Settings like this have been widely used in classroom video studies, aiming to gain a deeper understanding of the teacher-student interactions happening in classrooms and the factors that enhance students' learning outcomes (Janik & Seidel, 2009). In this dissertation, I explore conditions potentially enhancing our understanding of observers' assessments of teaching quality. To prepare the respective research questions, in the following, I provide a general overview of the promises and pitfalls that come with the observer perspective for the assessment of teaching quality (Chapter 1.2.1), before summarizing empirical findings about the psychometric quality of observer ratings (Chapter 1.2.2). Afterwards, I highlight the significance of rater training in classroom observation as central approach to ensure the quality of observer ratings and summarize the current state of research on rater training, leading to the research objective for the first paper of this dissertation (Chapter 1.2.3)

### 1.2.1 The Assessment of Teaching Quality Through Observation - Promises and Pitfalls

Observation is a fundamental method for studying complex interactions in social sciences (Ciesielska et al., 2018; Seidel & Prenzel, 2010). Observational methods can be classified as either participant or non-participant, and as either direct or indirect (Ciesielska et al., 2018). In participant observation, the observer actively engages within the observed group or organization, gaining an insider perspective, whereas non-participant observation involves no interaction, providing an outsider perspective. Direct observation refers to events observed live, in real-time, whereas indirect observation relies on previously collected data, such as video recordings or written descriptions of social interactions (Ciesielska et al., 2018). In the context of classroom observations, external observers usually take on a non-participant role. Both direct in-class observations and indirect observations using classroom videos are utilized (Curby et al., 2016; Jentsch et al., 2024); however, video-based observations have become the standard, as direct classroom observation is not always feasible and video-based methods offer promising opportunities for classroom research and teacher education (Janik & Seidel, 2009; Junker et al., 2022).

Assessing teaching quality from the observer perspective offers substantial potential for gaining in-depth insights into teacher-learner interactions in the classroom, benefiting both educational research and practice. The unique and important role observers hold in assessing teaching quality becomes visible with researchers describing them as the gold standard (Helmke, 2009) and "the most direct way to measure instructional quality" (Clare et al., 2001, p. 2). The judgments of external observers are regarded as especially valuable for several reasons. Unlike students, who may have a limited understanding of didactical or pedagogical principles (Clausen, 2002), observers are explicitly trained to recognize specific aspects of teaching quality and are expected to draw valid inferences from teachers' instructional practices and complex classroom interactions (Kunter & Baumert, 2007; Petko et al., 2003). Observers also typically observe a large number of lessons, which provides them with the opportunity to compare across diverse lessons (Clausen, 2002; Rakoczy, 2008). Additionally, observers are not directly involved in the teaching process, which is assumed to allow for a more objective assessment of quality (Petko et al., 2003; Pianta et al., 2008; Storms, 1973). In contrast, for example, self-assessments by involved teachers may be influenced by self-serving biases (Clausen, 2002).

Beside the potential of observers' assessments of teaching quality, there are central conceptual pitfalls associated with this perspective on teaching quality. One central pitfall of the observer perspective lies in the construct of teaching quality itself. Teaching quality, constituted by three basic dimensions, becomes visible in the interaction between teachers and students in the classroom. However, different aspects of teaching quality come with different levels of observability from an external perspective. For example, the basic dimension of classroom management is characterized by a well-organized classroom without major disruptions and an effectively used time on task (Kunter & Trautwein, 2013). These aspects of teaching quality can be observed from an external perspective quite easily. However, there are examples of quality-relevant teaching behavior, that is not easily available to an external observer (Fauth et al., 2020; Göllner et al., 2016). For example, the students' individual depth of processing the learning content, which is relevant to the basic dimension of cognitive activation, or the students' feeling of being emotionally supported by the teacher, which is relevant to the basic dimension of student support, are aspects of teaching quality that are not directly available to the observer but need to be assessed based on inferences the observer draws from the observed teacher- and student-behavior (Fauth et al., 2020; Vazire, 2010). At the same time, these aspects have been shown to be predictive of students' learning outcomes and,

therefore, are considered as important aspects of teaching quality (Baumert et al., 2010; Fauth et al., 2014; Göllner et al., 2018; Scherer et al., 2016; Wagner et al., 2016).

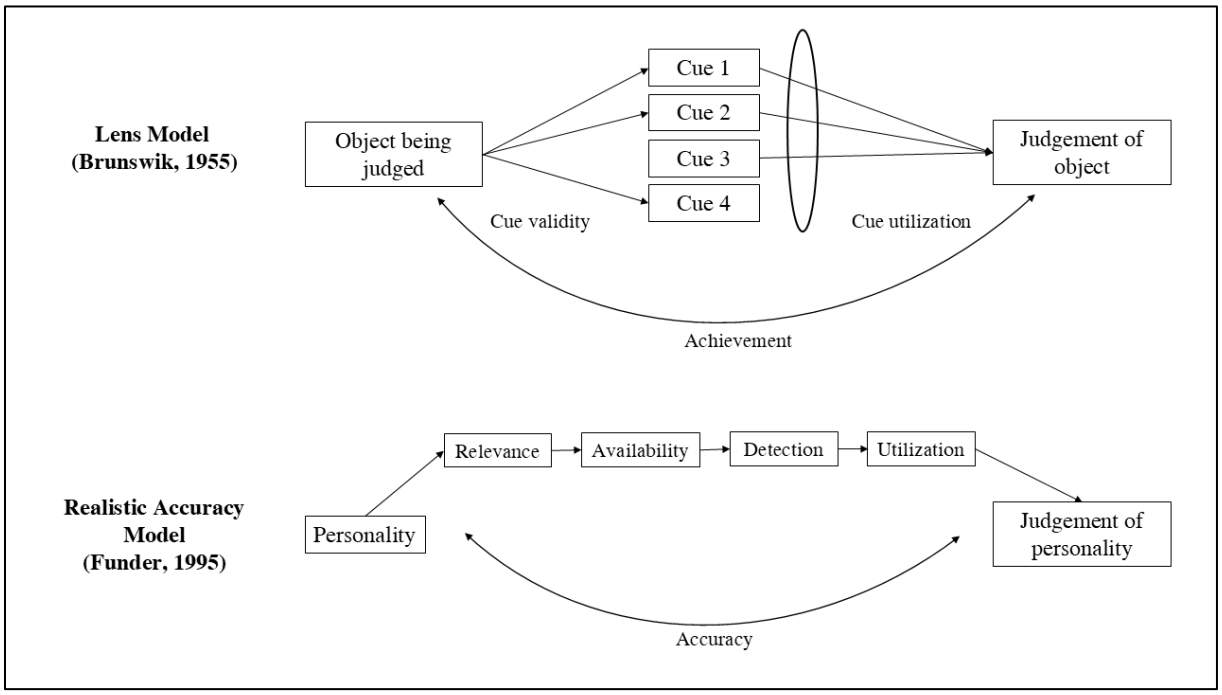
Another fundamental pitfall of the observer perspective lies in the observers as individuals themselves. Despite their seemingly objective outsider perspective on teaching, observers are never entirely free from idiosyncratic perceptions of the classroom environment. Fauth et al. (2020) state that „our knowledge of the world is and will always be an idiosyncratic construction that is fundamentally affected by our individual preconceptions and schemes of perception" (Fauth et al., 2020, p. 140). The proportion of ratings shaped not by the construct being assessed but by the idiosyncratic tendencies of the rater is referred to as rater bias (Myford & Wolfe, 2003, 2004; Wang & Engelhard, 2019). Rater bias describes the disagreements among raters arising from varying interpretations of rating scales or idiosyncratic perceptions of the evaluated subject (Hoyt, 2000). Consequently, rater bias is considered a source of measurement error (Praetorius et al., 2012). The literature on classroom observations has described different kinds of rater bias affecting the observation-based assessment of teaching quality. Prominent examples of rater bias are the effects of severity or leniency of raters, halo effects, central tendency, or restriction of range (for overviews, see Lenske, 2016; Praetorius, 2014).

The described conceptual pitfalls of the observer perspective on teaching quality are also reflected in scientific theories of judgement accuracy (Letzring & Funder, 2019). An early but established conceptual framework for rater-mediated assessments is Brunswik's (1955) *Lens Model* (Wang & Engelhard, 2019). Brunswik (1955) created the Lens Model of perceptual consistency to explain how raters make judgements of observed objects. His model assumes that we can never directly experience the world surrounding us but need to make judgements based on observable cues. These cues about the object being judged come with varying cue validity. In the process of judging the object, judges use these cues to draw inferences about the object. The judgements will be more accurate when cues with high cue validity are being used for the assessment (Brunswik, 1955). Based on Brunswik's Lens Model, the personality psychologist David Funder (1995) developed the Realistic-Accuracy Model (RAM) in the context of personality judgement. The RAM broke Brunswik's Lens Model down into a four-step process necessary for making accurate judgements of personality traits (Funder, 1995; Letzring & Funder, 2019). Behavioral cues need to be 1) relevant to the personality trait being assessed and 2) available to the rater in the external environment. Furthermore, behavioral cues need to be 3) detected by the rater and 4) utilized as cues for the assessment of personality. Figure 2 provides a graphical overview of the two models of judgement accuracy by Brunswik

(1955) and Funder (1995). In the Lens Model, the oval represents the lens through which the cues pass in the process of cue utilization. The visualization of the Lens Model shows that not all for judgement utilized cues are valid. In the RAM, cue validity corresponds with the relevance of the behavioral cues used for personality judgement. The RAM adds that cues must be available, and cue utilization is broken down into the steps of detection and utilization (Funder, 2001; Letzring & Funder, 2019).

**Figure 2**

*Overview of Two Models of Judgement Accuracy: Brunswik's Lenns Model and Funder's Realistic-Accuracy Model (adapted from Letzring & Funder, 2019)*



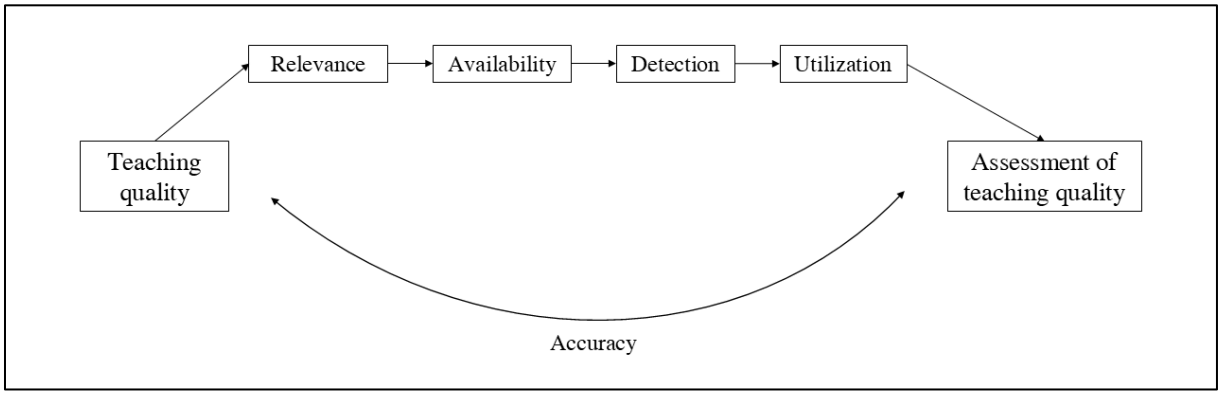

The RAM (Funder, 1995) offers a valuable theoretical framework that can be applied to assessing teaching quality through external observers. Funder's (1995) four-step process for personality judgement can be adapted into a parallel four-step process for observers to accurately assess teaching quality. Given its strong conceptual fit, I adapt Funder's RAM (1995) and propose a model of observers' assessment accuracy in classroom observations, designed to support the systematic investigation of the assessment of teaching quality through classroom observation.

Adapting the theoretical considerations of the RAM to the context of classroom observations, the behavioral cues from the context of personality judgement equal the teacher-student interactions observable in a classroom in the context of teaching-quality assessment.

For observers to provide an accurate assessment of teaching quality, the observed teacher- and student behaviors need to be 1) relevant for the quality of the teaching and 2) available to the observers from their outsider perspective. This second step contains the first central pitfall I described above, the limited observability of specific aspects of teaching quality, for example from the dimension of student support (Fauth et al., 2020; Vazire, 2010). Furthermore, the relevant and available teaching behavior needs to be 3) successfully detected (noticed) by the observers as well as finally 4) utilized for their assessment of teaching quality. The third and fourth steps rely on the observer and are, therefore, potentially affected by rater bias, the second central pitfall I described above. Figure 3 provides a graphical illustration of the RAM adapted to the context of teaching-quality assessment in classroom observations – the first step I take to propose a theoretical model of observers' assessment accuracy.

**Figure 3**

*Step 1 in Proposing a Model of Observers' Assessment Accuracy in Classroom Observations*




To sum up, external observers offer distinct advantages in assessing teaching quality compared to teachers and students directly involved in classroom teaching. However, the observer perspective also involves central pitfalls that may compromise the accuracy of their assessments. To address these limitations, the theoretical model of judgment accuracy by Funder (1995) can be adapted to the context of classroom observations, offering valuable insights into the processes underlying observers' assessments and enhancing our understanding of rating accuracy. To further refine the proposed model of observers' assessment accuracy in classroom observations, I examine additional factors influencing observation-based assessments of teaching quality.

### 1.2.2 Psychometric Quality of Observer Ratings of Teaching Quality

Observer ratings of teaching quality have been widely used in empirical studies, where observers provided ratings of teaching quality using standardized classroom observation systems (Bell et al., 2019; Praetorius & Charalambous, 2018), such as the CLASS (Pianta et al., 2008) or the Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2013). Standardized classroom observation systems are frameworks designed to systematically assess specific dimensions of teaching (Praetorius & Charalambous, 2018). Instead of asking observers for their general impressions of the teaching, these systems rely on protocols that specify teaching behaviors and instructional practices to be systematically observed and assessed on a numeric scale using rating items. To ensure consistency and reliability of ratings, these systems usually incorporate detailed observation manuals and rating guidelines that help observers interpret and evaluate classroom interactions with minimal idiosyncratic bias (Bell et al., 2019; Hill et al., 2012).

Over the years, empirical studies have paid increasing attention to the psychometric quality of observer ratings of teaching quality provided with standardized classroom observation systems. Even though those systems are created to standardize the assessment of observable aspects of teaching quality, it has become evident that the pitfalls of the observer perspective on teaching quality (Chapter 1.2.1) are also reflected in the empirical findings on the psychometric quality of observer ratings. Impactful work highlighting the limitations of observer ratings was provided by Praetorius (2014). In her work, Praetorius (2014) investigated how raters observed and assessed teaching quality in classroom videos. Her studies emphasized that despite using standardized classroom observation systems, observer ratings can suffer from issues of reliability and validity. Specifically, she found that 12-40% of the variance of ratings could be explained by rater bias, depending on the scale level and dimension of teaching quality (Praetorius et al., 2012). Furthermore, she found inconsistencies between raters in interpreting items or applying rating scales (Praetorius, 2014). Different studies using standardized classroom observation systems to assess teaching quality found comparable levels of variance explained by raters' idiosyncratic perceptions (e.g., Bell et al., 2014; Hill et al., 2012; Jones & Bergin, 2019; Lotz et al., 2013; Marder & Walkington, 2014). Across studies, it becomes clear that observers have fewer problems accurately observing and rating aspects regarding classroom management compared to teaching-quality dimensions like cognitive activation or student support (e.g., Bell et al., 2014; Bergin et al., 2017).

Due to the limited psychometric quality of observer ratings, researchers discussed the necessary number of lessons and number of raters to secure reliable ratings (Schlesinger & Jentsch, 2016). In general, the more lessons and raters are used to assess teaching quality, the higher the reliability of ratings (Ho & Kane, 2013; Praetorius et al., 2012). However, again there are differences depending on which dimension of teaching quality is assessed. Whereas, for example, one lesson was found to be sufficient to assess classroom management and student support, nine lessons were found to be relevant to assess cognitive activation (Praetorius et al., 2014). However, these results always rely on a specific observation system, and therefore, a direct transfer to other contexts is not always possible (Hill et al., 2012).

Regarding validity, researchers generated different validity arguments for using observer ratings of teaching quality (Bell et al., 2012; Hill et al., 2012; Liu et al., 2019; White, 2022). Examples of common validity arguments are a strong interrater agreement among raters (e.g., Martin-Raugh et al., 2016), factor structure of ratings (e.g., Li et al., 2020) or correlations of ratings with relevant external criteria, such as student achievement (e.g., Kane & Staiger, 2012). However, White and Klette (2023) make clear that the validity of observer ratings can vary widely across interpretations. This means that the validity of observer ratings must always be justified based on the intended interpretation and use of the information they provide (Bell & Gitomer, 2023; White & Klette, 2024).

To sum up, research highlights the psychometric limitations of observer ratings for assessing teaching quality. Given the importance of accurately assessing teaching quality for practice and research, various strategies have been developed to enhance the quality of observer ratings within specific classroom observation systems. These strategies include, for example, to recruit raters with a specific professional background, to have lessons scored by multiple raters (double-scoring), or to train and certify raters in the use of a specific classroom observation system (for an overview, see White & Ronfeldt, 2024). As rater training in classroom observation is a crucial part of this dissertation, I take a closer look at the strategy of training raters to ensure rating quality in the following.

### 1.2.3 Rater Training in Classroom Observations

Effective training is a critical prerequisite for producing consistent, reproducible, and accurate ratings across different assessment areas (Johnson et al., 2009). Also in the context of assessing teaching quality through classroom observations, training raters in using standardized classroom observation systems is one of the widely used quality control procedures for observer ratings (Bell et al., 2014, 2019; White & Klette, 2024; White & Ronfeldt, 2024).

Training raters in using a standardized classroom observation system falls under the category of frame-of-reference (FOR) training. Initially proposed by Bernardin and Buckley (1981), FOR training prepares raters to apply a shared conceptual framework (frame-of-reference) when observing and evaluating a subject (Roch et al., 2012). This type of training provides raters with detailed information about the dimensions of the target construct and the rating scales, along with positive and negative examples of the aspects being assessed (Gorman et al., 2015). This applies to typical rater training in classroom observations, where observers are trained to assess teaching quality using a standardized classroom observation system grounded in a specific conceptualization of teaching quality (frame-of-reference), for example the TBD.

In the United States, the influential Measures of Effective Teaching (MET) project explored how different assessment methods, including classroom observations, could best inform teachers about their teaching quality (e.g., Kane & Staiger, 2012). As part of the MET project, Joe et al. (2013) summarized practical foundations for effective classroom observations, including a description of rater training in classroom observation. The authors emphasize that for observers to deliver consistent and accurate ratings of teaching quality, they must share a common understanding of what constitutes each quality level across the different rating scales included in an observation system. Consequently, rater training in classroom observation aims to deepen observers' understanding of the system's dimensions and to provide them with targeted practice to apply the observation system accurately (Joe et al., 2013).

Rater training can be designed as face-to-face, online, or hybrid training. Typical training objectives include familiarizing raters with the observation system's framework for teaching quality and developing specific observation skills, such as distinguishing quality-relevant classroom events from less significant ones. This includes thoroughly familiarizing raters with the observation manual describing the rating items and procedures for the respective observation system (Joe et al., 2013). Working with the observation manual, observers learn how to accurately apply the system's rating scales. Furthermore, typical rater training makes observers aware of potential rater biases (Joe et al., 2013). Classroom videos are frequently used in rater training to illustrate positive and negative examples of teaching quality and to provide practice material for assessing teaching quality (Bell et al., 2019; Joe et al., 2013). After completing a rater training program, participants often undergo certification testing, where they demonstrate their proficiency by accurately rating classroom videos against master ratings,

which serve as true scores for comparison. These master ratings are typically determined by the developers of the observation system (Joe et al., 2013; White, 2018).

Summarizing empirical results on the effectiveness of rater training, I focus on the general effectiveness of FOR training across various contexts in the first step. The literature on FOR training effectiveness shows that practicing the assessment improves rating accuracy and that receiving feedback on the rating performance even increases this improvement (Elder et al., 2005; Ivancevich, 1979; Roch et al., 2012; Uggerslev & Sulsky, 2008). The meta-analysis by Roch et al. (2012), however, reveals that the effectiveness of FOR training depends on the operationalizations of accuracy. For example, FOR training proved to be especially effective in improving Borman's differential accuracy, which describes the degree to which ratings differentiate accurately between different dimensions of performance, as well as behavioral accuracy, which describes the precision with which raters observe, interpret and rate specific behaviors by the person being assessed (Borman, 1979). Regarding the effectiveness of rater training in the specific context of classroom observation, there is only limited empirical evidence on training effectiveness. Studies examining the effectiveness of rater training programs for ensuring the quality of observer ratings of teaching quality have primarily focused on assessing rating quality at the end of the training. They found that ratings mainly met the criteria of satisfaction after rater training, with variations of rating quality, depending on the specific aspect of teaching quality being assessed (Bell et al., 2014; Bergin et al., 2017; Cash et al., 2012). However, empirical research has not yet explored how rating quality develops over the course of training, how much training is necessary for raters to provide accurate observer ratings, or which training components are particularly effective for rating-quality development. Given the importance of rater training as a quality procedure for the assessment of teaching quality through classroom observation (Bell et al., 2014, 2019; White & Klette, 2024; White & Ronfeldt, 2024), it is essential to intensify research on rater training in classroom observation to design effective training programs. Addressing this research gap is the overarching objective of the first of the three studies included in this dissertation.

The significance of training for the success of observer ratings of teaching quality shows that the observers' disposition, including the level of training an observer has in providing observer ratings of teaching quality, is a crucial component for a model of observers' assessment accuracy in classroom observations. For this reason, I extend the adapted model in the second step with the observers' disposition influencing the process of teaching-quality assessment (see Figure 4).

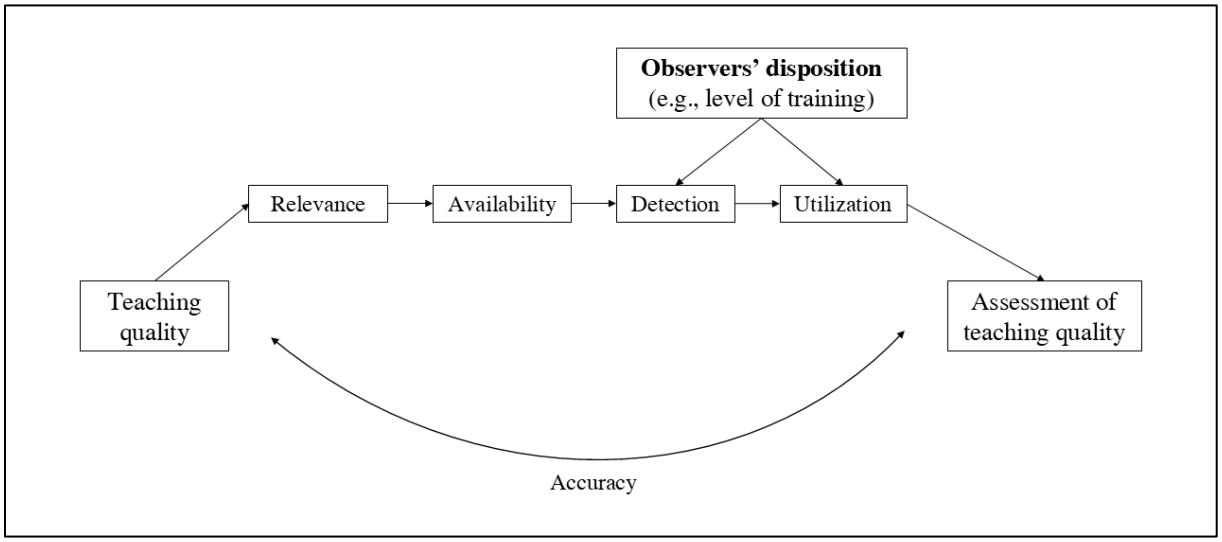
Two potential mechanisms by which the observers' disposition influences observer ratings can be considered:

1. Different event detection: The observers' disposition, for example the knowledge about the rating items and observable indicators of a specific classroom observation system, may shape their observation focus during classroom observation, impacting the detection of relevant classroom events. This difference in detection would subsequently affect the utilization of those events in teaching-quality assessments.
2. Different event utilization: Alternatively, even with the same events detected by an observer, the way these events are utilized may also be influenced directly, depending on the observers' dispositions. For example, observers might weigh a detected event as more or less important for assessing a specific aspect of teaching quality, based on their knowledge about how important this event is for the quality of teaching.

Given these two plausible mechanisms of how the observers' disposition affects classroom observations, the extended model includes arrows pointing from the observers' disposition to both the detection and utilization of quality-relevant classroom events.

**Figure 4**

*Step 2 in Proposing a Model of Observers' Assessment Accuracy in Classroom Observations*



## 1.3 Classroom Observation in Video-Based Environments

Throughout this dissertation, it has become evident that classroom observations are often conducted not in actual classrooms but through classroom videos. Video-based observations are used across various contexts in teacher education and professional development, where both pre-service and in-service teachers observe classroom videos to discuss teaching quality and reflect on ways to enhance their own practice (Gaudin & Chaliès, 2015; Santagata et al., 2005), and in educational research, where external raters observe and assess specific aspects of the classroom interaction (Janik & Seidel, 2009; Seidel & Thiel, 2017). Given the widespread use of video for conducting classroom observations, important questions arise: What role does the observation environment play in assessing teaching quality? Could the observation environment also influence the observers' assessment accuracy discussed in Chapter 1.2?

In this dissertation, I investigate the role of the classroom video environments for observer ratings of teaching quality, aiming to advance our understanding of these ratings. To establish the specific research objectives, the following chapter reviews current research on video-based classroom observation environments. I begin by examining the role of classroom videos as representations of teaching practice within teacher education and research (Chapter 1.3.1). Then, I introduce 360-degree videos as an emerging format that offers new possibilities for creating immersive classroom observation environments (Chapter 1.3.2). Finally, I summarize existing knowledge on how different classroom video environments impact classroom observations and the resulting ratings of teaching quality (Chapter 1.3.3). This discussion leads directly to the research objectives addressed in my dissertation's second and third studies.

### 1.3.1 Classroom Videos as Representations of Teaching Practice

Observing classroom teaching is invaluable for analyzing and learning from classroom interactions in teacher education and research. However, conducting classroom observations in real-time in the actual classroom can be challenging. Because teaching is a highly complex process (Praetorius & Charalambous, 2023), live observations do not allow for systematic and detailed analysis (Seidel, 2022). For this reason, researchers and teacher educators have increasingly turned to collecting video data, using classroom videos as representations of teaching practice (Brophy, 2004; Gaudin & Chaliès, 2015; Grossman, 2021; Seidel & Thiel, 2017). Classroom videos offer numerous advantages. As Klette (2016) notes, they enable "more

precise, complete, and subtle analyses of teaching/learning processes" (Klette, 2016, p. 1) compared to live observations. These detailed analyses are possible because classroom videos are able to capture the classroom environment from multiple perspectives, allow a focused examination of individual interactions, can be replayed multiple times, and thus enable observers to analyze and discuss teaching and learning in detail (Janik & Seidel, 2009; Junker et al., 2022; Krammer & Reusser, 2005; Syring et al., 2015).

Classroom videos come in various forms, differing in aspects such as authenticity, length, perspective on classroom interactions, and production methods. In terms of authenticity, some videos capture naturally occurring classroom activities, providing an unaltered view of teaching and learning. In contrast, staged videos use scripted teacher-student interactions to highlight specific, often rare, teaching events or didactically relevant practices (Codreanu et al., 2020; Piwowar et al., 2018; Seidel et al., 2022). In terms of length, videos range from short, only second-long video sequences, so-called "thin slices" (e.g., Begrich et al., 2021), to minute-long classroom video sequences, to videos from entire lessons or even teaching units (e.g., Klieme et al., 2009). When it comes to perspective, some videos emphasize the teacher, others focus on students, while others capture whole-class interactions, often reflected in camera angles (Kilburn, 2014; Otrel-Cass et al., 2010). Additionally, the way videos are produced can vary substantially. Some videos are professionally produced as part of large-scale studies (e.g., Ainley & Carstens, 2018), while others are created on a smaller scale, with teachers recording their own lessons for self-reflection or peer discussion (Seidel et al., 2011; Tripp & Rich, 2012). Choosing the appropriate type of video depends on the specific objective (for an overview, see Gaudin & Chaliès, 2015).

Classroom videos represent teaching practices across diverse contexts. In teacher education, they help bridge the theory-practice gap by allowing pre-service teachers to connect theoretical concepts with practical illustrations of teaching (Korthagen, 2007; McGarr et al., 2017). In this way, working with videos can prepare pre-service teachers for their own future classrooms. The main objectives of using videos in teacher education are to develop skills in interpreting and reflecting on classroom scenarios—central to what is termed teachers' professional vision (Seidel & Stürmer, 2014)—and to practice responses to various situations they may encounter as teachers (Gaudin & Chaliès, 2015). Research shows that working with classroom videos effectively promotes teacher competencies related to teaching quality, such as professional vision of classroom interactions (Sherin & van Es, 2005; Weber et al., 2018),

feedback competence (Muñiz-Rodríguez et al., 2018; Prilop et al., 2020), and reflective skills (Hamel & Viau-Guay, 2019; Shek et al., 2021).

In educational research, classroom videos have been widely used in the context of video studies, such as TIMSS (Hiebert & Stigler, 2000), the Teaching and Learning International Survey (TALIS; Ainley & Carstens, 2018), the German-Swiss Pythagoras study (Klieme et al., 2009) or the Nordic Linking Instruction and Student Achievement (LISA) study (Sigurjónsson et al., 2022; Tengberg et al., 2022). In such studies, researchers typically analyze the teaching captured on video in detail, often by rating teaching quality using standardized classroom observation systems (Bell et al., 2019; Klette, 2023; Praetorius & Charalambous, 2018). Over recent decades, classroom video research has substantially advanced our understanding of teaching and learning processes (Klette, 2023). For example, systematic video analyses have contributed to developing and validating frameworks for conceptualizing teaching quality, such as the TBD (Klieme et al., 2009; Praetorius et al., 2018). Furthermore, classroom video studies made it possible to systematically compare teaching quality across countries (White & Klette, 2023) or have highlighted the impact of core aspects of teaching quality on students' learning outcomes (e. g., Fauth et al., 2014; Lipowsky et al., 2009).

It becomes evident that classroom videos are a unique and valuable tool for making classroom interactions observable and analyzable in both teacher education and educational research (Klette, 2016; Seidel, 2022). However, traditional approaches to using classroom videos on a computer screen also have limitations. One central limitation is that the complex interactions depicted in classroom videos can be overwhelming for observers, especially for novice teachers (Erickson, 2007; Gaudin & Chalies, 2015). Therefore, how videos are utilized is critical to achieving the intended learning goals. For example, a well-established strategy for reducing cognitive load during video observation is to use short video sequences with explicit prompts to guide focus (Brunvand, 2010; Wilkes et al., 2022). Another significant limitation of traditional classroom videos is that they represent only one aspect of classroom reality, as they are filmed from a specific camera angle. This "keyhole effect" (van Es & Sherin, 2002) restricts observers from noticing multiple events that often occur simultaneously in classrooms. Observing the entire classroom interaction without a restricted angle of view could allow for a more comprehensive detection of quality-relevant interactions between teachers and students. One way to address the limitation of the restricted field of view due to the camera angle is by using 360-degree video technology (e.g., Evens et al., 2023). Since classroom observations with

360-degree video are a central part of this dissertation, I will introduce this video type in the following chapter.

### 1.3.2 Increasing Immersion: 360-Degree Classroom Videos

In recent years, digital innovations have expanded opportunities for teaching and learning (Scheiter, 2021). This is also true for the use of videos as research and learning tool, where technological advancements have broadened possibilities for recording and presenting video content. A major development in this area is using 360-degree videos, also known as immersive or spherical videos (Evens et al., 2023; Ranieri et al., 2022; Snelson & Hsu, 2020). 360-degree videos are captured with omnidirectional or multi-camera systems, recording in all directions simultaneously to create a full spherical field of view. This allows video observation without a fixed camera angle, giving users the freedom to choose their perspective within the video environment. 360-degree videos can be viewed on various devices, ranging from everyday items like smartphones, tablets, and computers to more sophisticated equipment, such as HMDs or large cylindrical projection systems that fully surround viewers (Rupp et al., 2019). Depending on the device, users can select their viewing angle by scrolling, clicking, or moving their heads (Graham et al., 2023; Kavanagh et al., 2016; Snelson & Hsu, 2020).

Compared to traditional videos, 360-degree videos provide a more immersive experience of the video content (e.g., Snelson & Hsu, 2020). Immersion, often discussed as a technology-oriented aspect of VR, refers to the extent to which technology creates a vivid, surrounding, and lifelike illusion of a virtual environment for the viewer (Slater & Wilbur, 1997). The feeling of being present within an environment is described as a psychological, perceptual, and cognitive consequence of immersion (Calvert & Abadia, 2020; Slater & Wilbur, 1997). Different devices for displaying 360-degree videos offer varying levels of immersion: the more an observer is isolated from the external world and surrounded by the video environment, the greater the immersive effect the device induces (Rupp et al., 2019).

360-degree videos are increasingly used in educational settings (Atal et al., 2023; Evens et al., 2023; Ranieri et al., 2022; Rosendahl & Wagner, 2023; Snelson & Hsu, 2020). Literature reviews on 360-degree videos in education reveal that these videos are predominantly applied in higher education, particularly in healthcare and teacher education (Evens et al., 2023). In the context of teacher education, 360-degree classroom videos are used to support the development of professional competencies, primarily focusing on teachers' noticing and reflection skills (Atal et al., 2023). Rosendahl and Wagner (2023) identify five key advantages over traditional videos: increased motivation and interest, authentic learning scenarios, immersive learning

experiences, multi-perspective observation, and individualized learning (Rosendahl & Wagner, 2023). Research highlights positive effects of 360-degree videos on learning outcomes, including enhanced declarative, procedural, and skill-based knowledge (Evens et al., 2023; Snelson & Hsu, 2020). Furthermore, 360-degree videos enhance feelings of immersion and presence, self-efficacy, empathy, reflection, engagement, satisfaction, and enjoyment (Evens et al., 2023; Snelson & Hsu, 2020). However, the literature acknowledges limitations, such as the risk of cognitive overload from the 360-degree environment's complexity (Parong & Mayer, 2021; Roche et al., 2021; Sweller, 2011) and the possibility of users overlooking relevant events due to the unrestricted field-of-view (e.g., Ardisara & Fung, 2018).

Overall, 360-degree classroom videos hold promising opportunities, with initial applications already emerging in teacher education to support the development of professional competencies. However, they also present challenges, such as potential cognitive overload and the risk of missing critical events within the immersive environment. Given these changed characteristics compared to traditional videos, it becomes clear that 360-degree videos create a different observation environment for classroom observations. For the assessment of teaching quality through classroom observation, this brings up the question of how 360-degree video environments impact observers' classroom observations, their perception of quality-relevant teaching events and, consequently, the resulting observer ratings of teaching quality.

### 1.3.3 How Observation Environments Impact Classroom Observations

In teacher education and research, classroom observation often takes place using classroom videos as representations of teaching practice. The design of video-based observation environments can vary significantly, depending on factors such as camera perspectives, video type, and display device. For instance, observing classroom teaching from two different camera angles in a 16:9 video format on a computer may differ substantially from observing the same lesson from a viewpoint within the classroom in a 360-degree video displayed through VR glasses. In this chapter, I review research on how different observation environments affect classroom observations and observer ratings of teaching quality.

When examining differences between observation environments, the first question to arise is how classroom observation and ratings of teaching quality differ between live and video observations. Some studies have compared observer ratings of teaching quality across these modes (Casabianca et al., 2013; Curby et al., 2016; Jentsch et al., 2024). In terms of absolute ratings, all three studies reported differences, though the direction varied. Casabianca et al. (2013) found that video ratings were slightly higher, whereas Curby et al. (2016) reported

marginally lower ratings for video observations. Jentsch et al. (2024) observed a nuanced pattern depending on the basic dimension of teaching quality assessed: classroom management was rated lower in video observations, while cognitive activation received higher ratings. The authors attribute these differences to the characteristics of each observation mode, such as differences in what observers can see and hear accurately, for example regarding students' discourse (Jentsch et al., 2024). Despite these variations, the ranking of the lessons' quality was consistent across observation modes (Casabianca et al., 2013; Jentsch et al., 2024). Regarding the reliability of ratings, both Curby et al. (2016) and Jentsch et al. (2024) found acceptable and mostly comparable reliabilities across modes, except for classroom management, which was rated more reliably in live observations (Jentsch et al., 2024). Casabianca et al. (2013) found variations in how rating scales were utilized across observation modes, resulting in differences in reliability and the conclusions drawn from individual lessons. Curby et al. (2016) provided evidence of predictive validity for student outcomes in both modes, concluding that neither mode is uniformly superior. In summary, research comparing live and video classroom observations indicates that both approaches are viable for assessing teaching quality. However, certain aspects of teaching quality may be more effectively captured in one mode than the other. For instance, aspects of classroom climate or students' discourse might be better assessed when an observer is physically present in the classroom (Jentsch et al., 2024).

Within the research field on video-based classroom observations, several studies have investigated differences between different forms of video-based observation environments. One line of research is to investigate differences when the same teaching is observed from different camera perspectives (Cortina et al., 2018; Paulicke et al., 2019; Wyss et al., 2023). The studies by Cortina et al. (2018) and Wyss et al. (2023) both compared what observers noticed and how they reflected on classroom videos between different camera perspectives (see also concept of teachers' professional vision; Seidel & Stürmer, 2014). Cortina et al. (2018) used recordings from the teacher's field of view and recordings from the students' field of view and found more student-focused comments about the teaching for the video recordings from the teacher's perspective. Wyss et al. (2023) compared three different camera perspectives (back camera, front camera, eye-tracking camera from teacher's field of view) regarding observers' visual attention on the teacher, the students and the learning material in classroom videos as well as verbal reports on the observed teaching. They found that, depending on the perspective, different objects and individuals are in the observers' visual focus of attention, but verbal reflections on the teaching were not affected by the camera perspective (Wyss et al., 2023). The study by Paulicke et al. (2019) investigated differences in observer ratings of teaching quality

for the same lessons recorded from three different camera perspectives (camera from the students' perspective, camera from teacher's perspective, and overview camera). They found significantly differing teaching-quality ratings for the different camera perspectives for all assessed dimensions of teaching quality. Furthermore, student-individual behavior was perceived primarily from the camera angle from the students' perspective (Paulicke et al., 2019). In summary, research comparing classroom observation and ratings of teaching quality from different camera perspectives makes clear that the observation environment created through camera perspective can result in substantial differences in how teaching is perceived, and teaching quality is rated.

Another line of research is to investigate the impact of video type on classroom observations, as 360-degree videos offer more immersive ways to experience classroom videos (see Chapter 1.3.2). First research has examined the effects of different video environments on teachers' classroom observation experiences and teaching assessments. Regarding observation experiences, studies found that observers feel more immersed and physically present in the observation environment in 360-degree videos (Ferdig & Kosko, 2020; Gold & Windscheid, 2020; Kunz & Zinn, 2022). Only one study on 360-degree classroom videos included observer ratings of teaching quality and found no significant differences in ratings between traditional classroom videos and 360-degree videos, whereas both video types were displayed on a computer screen (Gold & Windscheid, 2020). Different studies compared observers' noticing of critical classroom events between video types (Kosko et al., 2021, 2022). Kosko et al. (2021) compared pre-service teachers' noticing of critical events in mathematics lessons between traditional videos on computers, 360-degree videos on computers, and 360-degree videos on VR glasses. They found that the 360-degree VR condition resulted in pre-service teachers noticing critical events in more detail and focusing more on students' actions than in the other conditions. Kosko et al. (2022) investigated conditions of successful noticing within 360-degree VR videos. They found that the noticing of critical classroom events depends on how observers put students and the teacher in their field of view. In summary, research comparing classroom observation in different video types provides first evidence that there are large differences in the classroom observation experience between video types. Especially the feeling of being immersed and present in the classroom in immersive 360-degree video environments promises a more realistic observation compared to traditional video environments. Furthermore, first evidence on facilitated noticing of classroom events suggests that the observation environment might also impact ratings of teaching quality, but further research is needed to understand rating differences between video types in more depth. Gaining deeper insights into the effects of

immersive 360-degree video environment observation on classroom observations and ratings of teaching quality is the overarching objective of the second of the three studies included in this dissertation project.

The review of the literature on how different observation environments influence classroom observation reveals a critical insight: the observation environment matters for assessing teaching quality. To account for this, I extend the adapted model of observers' assessment accuracy in classroom observations in the third step by incorporating the component of the observation environment in which the observation occurs (see Figure 5).

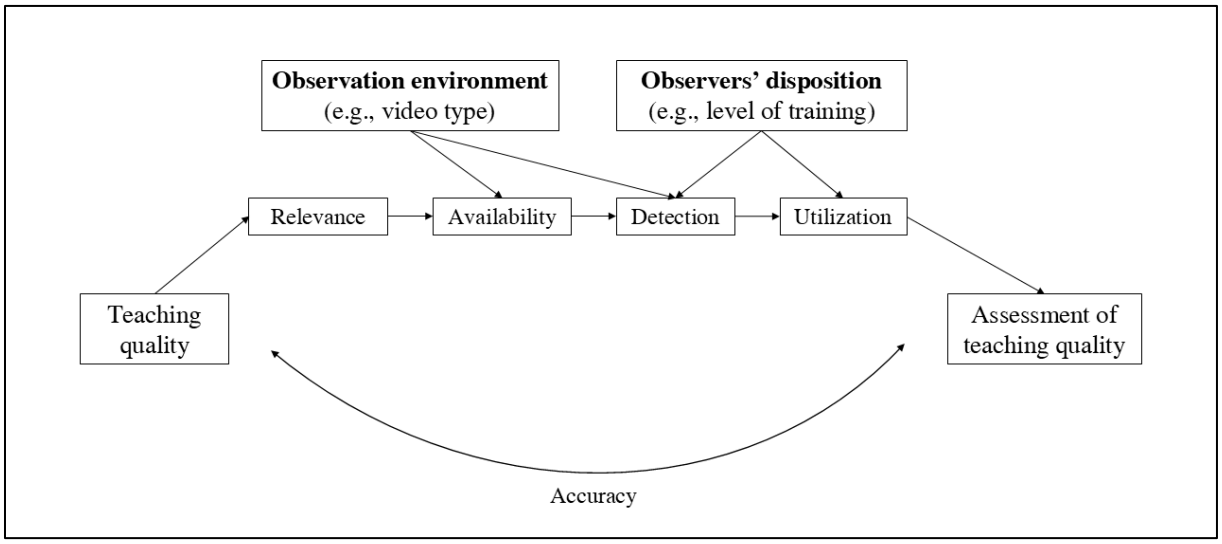
Building on the findings that different observation environments lead to a different noticing of classroom events, two potential mechanisms may explain how the observation environment impacts observer ratings.

1. Different availability of classroom events: Specific characteristics of the observation environment (for example the immersive nature of 360-degree videos, with their ability to present events from all angles around the observer) might lead to a different availability of classroom events for observation. These differences in availability could then influence the subsequent detection and utilization of these events when assessing teaching quality.
2. Different event detection: Alternatively, even with the same level of event availability across environments, the observation environment may influence the likelihood of observers detecting relevant events. This difference in detection probabilities would subsequently affect the utilization of those events in teaching-quality assessments.

Given these two plausible mechanisms for how the observation environment affects classroom observations, the extended model includes arrows pointing from the observation environment to both the availability and detection of quality-relevant classroom events. This extension acknowledges the complex interplay between the observation environment and the processes involved in classroom observation.

**Figure 5**

*Step 3 in Proposing a Model of Observers' Assessment Accuracy in Classroom Observations*




Considering the effects of the observation environment for the assessment of teaching quality in classroom observations, it is particularly relevant to see how a specific observation environment impacts the necessary steps of availability, detection and utilization of critical teaching events for the assessment of teaching quality. To investigate this interplay, increasing use should be made of process data of video-based classroom observation to understand not only the outcomes but also the processes behind classroom observations and ratings of teaching quality. In the following chapter, I highlight the importance of eye-tracking technology as one possible way to generate process data of classroom observations in video-based environments.

### 1.3.4 Using Eye-Tracking as Process Data of Classroom Observations

Classroom observation in video-based environments holds promising opportunities. One of these opportunities for educational research is that video observations allow the creation of a standardized research environment, facilitating the collection of process data of classroom observations, for example through eye-tracking technology (Grub et al., 2020; Just & Carpenter, 1976; Keskin et al., 2024). Eye-tracking is a technology used to measure where and for how long a person's gaze focuses, tracking eye movements using specialized devices such as cameras or infrared sensors (Holmqvist et al., 2011). Based on the so-called *eye-mind link* (Just & Carpenter, 1976), suggesting that gaze directly reflects cognitive processes, eye-tracking is used to provide insights into a person's visual attention or cognitive processing (Holmqvist et al., 2011). With reference to the context of assessing teaching quality through classroom

observations, eye-tracking data can provide researchers with deeper insights into visual attention and cognitive processes underlying classroom observation.

In eye-tracking research, several key indicators are used to analyze visual attention and cognitive processes. Fixation duration and fixation count are commonly employed to measure how long and how often a person looks at a particular area of interest (AOI), providing insights into attention allocation (Holmqvist et al., 2011). Saccades, or rapid eye movements between fixations, are analyzed to study search patterns and visual exploration. The length and direction of saccades can indicate how efficiently information is being processed or scanned (Rayner, 2009). Pupil diameter is another indicator, often associated with mental effort or emotional arousal; larger pupil sizes typically reflect higher mental workload or engagement (Laeng et al., 2012). Additionally, scan path analysis, which examines the sequence of eye movements, can reveal strategies in visual search and information gathering (Andrews & Coppola, 1999). These gaze-based indicators collected through eye-tracking technology allow researchers to infer underlying processes related to perception, attention, and decision-making in various contexts. For this reason, eye-tracking technology has been widely applied in fields like marketing, psychology, and education sciences (Holmqvist et al., 2011).

In classroom-video research, eye-tracking has been applied mostly in research on teacher professional vision (Grub et al., 2020; Keskin et al., 2024). Eye-tracking studies on teacher professional vision often focuses on how (pre-service) teachers observe, interpret, and make decisions based on classroom interactions, with eye-tracking providing objective data on where teachers direct their attention during critical teaching events. For example, studies have shown that expert teachers distribute their gaze more evenly across the classroom and tend to fixate more on students and their behaviors, while novice teachers focus more on instructional materials or less relevant areas in the classroom (Grub et al., 2020; Keskin et al., 2024).

Whereas eye-tracking research using classroom video observation has yielded valuable insights into teachers' professional vision, significant research gaps remain. First, such research has predominantly focused on traditional video environments displayed on standard computer screens, with the notable exception of a recent study by Kosko et al. (2024). However, immersive 360-degree videos also enable the tracking of observers' gaze through eye-tracking technology (Adhanom et al., 2023). Comparing gaze behavior during classroom observations in traditional versus immersive 360-degree video environments could provide valuable insights into the suitability of each video format for teacher education and research. Second, to date, no studies have applied eye-tracking to investigate how observers assess teaching quality in

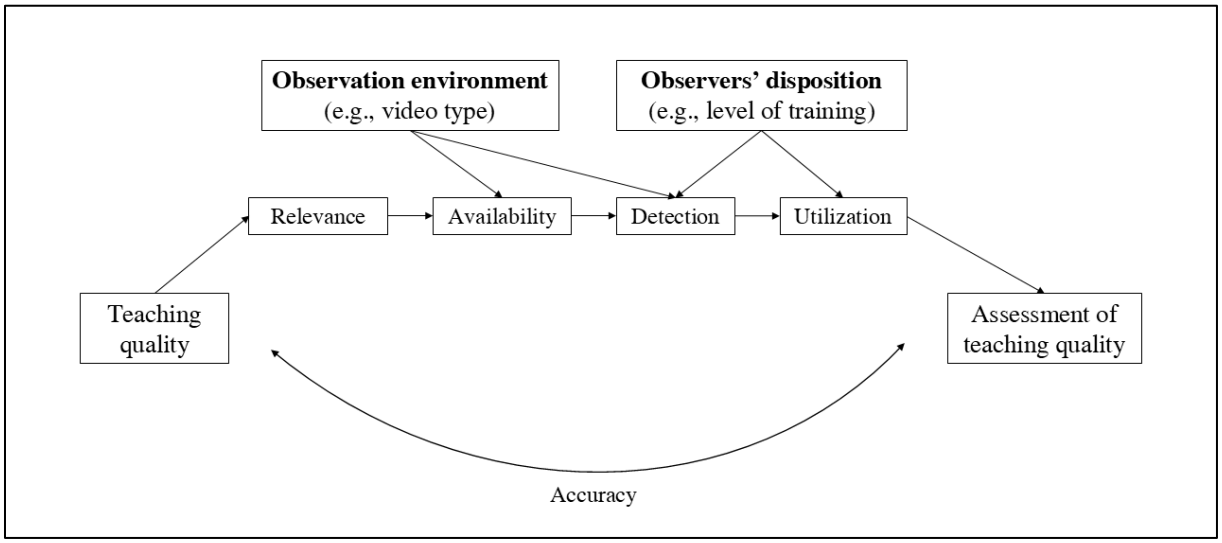
classroom videos. Research on observer ratings of teaching quality often emphasizes the ratings themselves as outcomes of classroom observations, rather than the observation process leading to those ratings. Integrating eye-tracking data as process data of classroom observations offers a promising opportunity to uncover the cognitive mechanisms underlying observation-based assessments of teaching quality. In the context of the proposed model of observers' assessment accuracy in classroom observations, tracking observers' gaze behavior during observations could provide objective data on the detection of quality-relevant classroom events—an essential step for accurate teaching-quality assessments in the model. Addressing this research gap and examining how observers' gaze behavior relates to their ratings of teaching across different classroom video environments is the overarching objective of the third study included in this dissertation

## 1.4 Model of Observers' Assessment Accuracy in Classroom Observations

Throughout this dissertation's introduction, I progressively adapted and extended Funder's (1995) RAM from the context of personality judgment to the context of assessing teaching quality through classroom observation. The resulting proposed model of observers' assessment accuracy in classroom observations is illustrated in Figure 6.

**Figure 6**

*Proposed Model of Observers' Assessment Accuracy in Classroom Observations*




This model explains the accuracy of observation-based assessments of teaching quality through a four-step process that describes the key stages of the classroom observation assessment process. In this chapter, I summarize the central theoretical assumptions underlying the proposed model.

1. **Relevance:** For observers to accurately assess teaching quality, the behavioral cues presented by the teacher and students in the teaching situation must be relevant to the assessment of teaching quality. For instance, the color of the teacher's shirt is irrelevant to the quality of teaching, whereas the formulation of the teacher's questions likely is quality-relevant. Without teacher-student interactions that are relevant to teaching quality, an accurate assessment of teaching quality is not possible. The relevance step can be described as a more situational step of the observation process (Funder, 1995).

2. **Availability:** For observers to accurately assess teaching quality, the quality-relevant behavioral cues exhibited by the teacher and students must be available from the observer perspective. For example, cues such as whether students are attentively listening to the teacher or are mind-wandering may not be directly observable. However, observable behaviors that reflect students' cognitive engagement, such as hand-raising (Böheim et al., 2020), may be available. Without relevant teacher-student interactions that are available to the observer, an accurate assessment of teaching quality is not possible. The availability step can also be described as a more situational step of the observation process (Funder, 1995).
3. **Detection:** For observers to accurately assess teaching quality, they must detect the quality-relevant and available behavioral cues exhibited by the teacher and students. For instance, specific student disruptions may be detected by the observer, but they can also be overlooked. When relevant and available teacher-student interactions are missed, an accurate assessment of the respective aspects of teaching quality becomes impossible. The detection step can be described as a more observer-specific step of the observation process (Funder, 1995).
4. **Utilization:** For observers to accurately assess teaching quality, the relevant and available behavioral cues they detected must also be effectively utilized in the assessment process. For example, if an observer notices that the teacher only asks superficial questions but fails to consider this as evidence of poor teaching quality in terms of asking challenging questions, the assessment will be inaccurate. When relevant and available teacher-student interactions are detected but not appropriately utilized, an accurate assessment of the respective aspects of teaching quality becomes impossible. The utilization step can also be described as a more observer-specific step of the observation process (Funder, 1995).

Besides this four-step process, the proposed model describes two central aspects that influence the assessment process:

1. **The observation environment:** The environment in which a classroom observation takes place (e.g., in the real classroom, with traditional classroom videos, or immersive 360-degree classroom videos) significantly impacts the assessment of teaching quality. Specifically, the observation environment may influence both the availability of information available to the observer and the likelihood of detecting this information.

2. **The observers' disposition:** The observers' disposition (e.g., their professional knowledge, subject-specific background, or level of training in using a standardized classroom observation system) significantly impacts the assessment of teaching quality. Specifically, the observers' disposition may influence both the detection and utilization of relevant, available behavioral cues, as detection and utilization are the more observer-specific steps of the classroom observation assessment process.

# 2

# AIMS AND RESEARCH QUESTIONS

# 2 Aims and Research Questions

As an important predictor of students' learning outcomes, teaching quality and how to assess it properly is a central issue in educational research and practice (Chapter 1.1). In this context, observing teaching quality from an outside perspective offers promising opportunities, but at the same time the observer perspective on teaching quality comes with several pitfalls and often results in ratings with limited psychometric quality (Chapter 1.2.1, Chapter 1.2.2). To support observers in their rating quality, rater training has been established as a quality procedure for observer ratings of teaching quality provided with standardized classroom observation systems (Chapter 1.2.3). In the context of rater training, but also in applications of classroom observations in teacher education and educational research, classroom videos have been the most common way to present the teaching to be observed and assessed (Chapter 1.3.1). Regarding classroom videos, technological advancements make new forms of video observations possible, for example through immersive 360-degree videos (Chapter 1.3.2). However, the observation environment impacts classroom observations and the resulting ratings of teaching quality (Chapter 1.3.3), which makes it necessary to investigate the advantages and disadvantages of different video environments to assess teaching quality through classroom observation. Thereby, including process data, such as eye-tracking data, holds promising opportunities to gain in-depths knowledge about cognitive processes behind the observation-based assessment of teaching quality (Chapter 1.3.4).

In the introduction of this dissertation, I have adapted Funder's (1995) RAM from the context of personality judgement to the context of teaching-quality assessment in classroom observations and proposed a model of observers' assessment accuracy in classroom observations (Chapter 1.4). Throughout the introduction, I have stepwise expanded this model by the observers' disposition and the observation environment as elements impacting the classroom observation assessment process. Consequently, the observers' disposition and the observation environment might also impact the accuracy of observer ratings of teaching quality. Based on the model I am proposing, the present dissertation aims to understand the conditions for observers to provide accurate assessments of teaching quality in (video-based) classroom observations more deeply. To achieve this aim, I present three empirical studies, addressing research gaps I identified in the literature. Each of the three studies takes a different approach to investigate conditions for accurate observer ratings of teaching quality. The three studies contribute to answering three overarching research questions:

1. How does rater training in classroom observation impact observer ratings of teaching quality?
2. How do different video environments impact classroom observations and observer ratings of teaching quality?
3. What insights can eye-tracking data from classroom observations provide about classroom observations and observer ratings of teaching quality?

In the following, I present the three empirical studies tackling these overarching research questions:

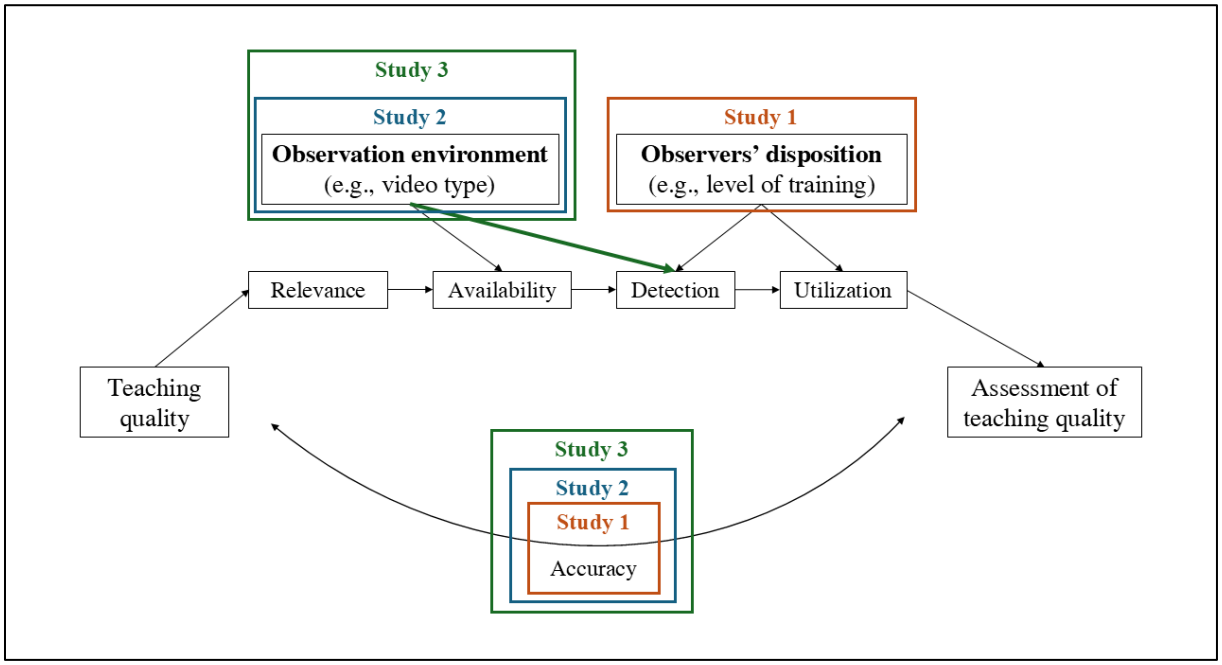
Study 1 (*Can teachers be trained to provide valid teaching-quality ratings?*) systematically investigated observer ratings of teaching quality throughout a rater training for in-service teachers. The participating teachers were trained in using a newly developed standardized classroom observation system for teaching-quality development in school practice in Baden-Württemberg, Germany (Fauth et al., 2021). The study focuses on investigating the reliability and validity of observer ratings of teaching quality. The reliability of ratings was examined through the interrater agreement between the teachers at five time points over the course of the rater training. Furthermore, three different validity arguments for the use of the teaching-quality ratings by the trained teachers were derived using three different approaches to validity: agreement with expert ratings, factor structure of ratings, and convergent correlations of ratings with an observation system assessing the same or similar aspects of teaching quality. Study 1 thereby focuses on the level of training as one aspect of the observers' disposition for providing accurate assessments of teaching quality. For this reason, Study 1 addresses Research Question 1 of this dissertation.

Study 2 (*Immersive insights: Unveiling the impact of 360-degree videos on pre-service teachers' classroom observation experiences and teaching-quality ratings*) compared pre-service teachers' classroom observation and observer ratings of teaching quality in two different video environments: traditional classroom videos, displayed on a computer and immersive 360-degree classroom videos displayed using VR glasses. Specifically, the study used self-reported cognitive, affective, and physiological observation experiences as well as absolute teaching-quality ratings to compare classroom observation and assessments of teaching quality between the video environments. For this reason, Study 2 addresses Research Question 2 of this dissertation, focusing on the outcomes of classroom observation in different video environments.

Study 3 (*Connecting gaze behavior and ratings of teaching quality*) examined observers' perceptual processes during classroom observations, using eye-tracking data as process data of classroom observations. Specifically, the study used gaze-based indicators relevant for assessing teaching quality (e.g., visual attention on relevant AOIs or pupil diameter) during previously defined critical events, which determined the quality of teaching in classroom videos. The study investigated associations between these gaze-based indicators during the observation of critical events and the accuracy of the resulting observer ratings of teaching quality, operationalized through the deviation from master ratings provided by experts. These associations between gaze behavior and rating accuracy were investigated in both traditional and immersive 360-degree video environments. For this reason, Study 3 addresses Research Question 3, but also contributes to Research Question 2 from a different point of view than Study 2. The three studies of my dissertation can be located in the proposed model of observers' assessment accuracy (see Figure 7).

**Figure 7**

*Location of the Empirical Studies in the Proposed Model of Observers' Assessment Accuracy in Classroom Observations*



# 3

# STUDY 1

Daltoè, T., Blank, J. L., Ruth-Herbein, E., Jaekel, A.-K., Göllner, R., & Fauth, B. (2024). *Can teachers be trained to provide valid teaching-quality ratings?* Manuscript submitted for publication.

**Abstract**

Teaching quality is a crucial determinant of students' academic achievement and motivation, and substantial research has deepened our understanding of its conceptualization and assessment. However, translating this theoretical knowledge into effective strategies that empower teachers to improve their practice remains a challenge. Observation-based feedback provided by trained teachers represents a promising approach to fostering teaching quality, yet concerns about the reliability of these ratings persist. This study explores a rater training program designed to enable teachers to serve as classroom observers who deliver formative feedback on teaching quality using a standardized classroom observation system. Ten in-service teachers participated in the training, rating classroom videos at multiple time points—before, during, and after the training. Our findings indicate that the training increased interrater agreement in teaching-quality ratings. We found validity evidence for the ratings of trained teachers, including alignment with expert ratings, factor structure of ratings, and convergent correlations with a comparable classroom observation system. These results underscore the potential for trained teachers to generate valid feedback on teaching quality, while also identifying specific aspects of teaching quality that warrant further attention in rater trainings.

## Introduction

For many years, teaching quality has been studied as a central determinant of students' academic achievement and motivation (e.g., Burroughs et al., 2019; Hattie, 2009). Considerable research has advanced our understanding of how teaching quality can be conceptualized (e.g., Panayiotou et al., 2021; Pianta & Hamre, 2009) and how it can be measured in a reliable and valid way (e.g., Senden et al., 2023; Wemmer-Rogh et al., 2024). However, there remains a pressing need to translate this theoretical knowledge about teaching quality into practical strategies that support teachers in developing their teaching quality (Bell & Gitomer, 2023).

One promising approach to support teaching-quality development in school practice is classroom observation (Gitomer, 2021; Martinez et al., 2016; Praetorius & Charalambous, 2018). In classroom observations, external raters observe teaching and provide assessments of teaching quality, often using standardized classroom observation systems (Bell et al., 2019; Gitomer, 2021). Previous research has often emphasized the limited psychometric quality of observer ratings of teaching quality (e.g., Kelly et al., 2020; Praetorius et al., 2012; White & Klette, 2024). Beyond examining the psychometric quality of ratings, it is equally important to explore if they can provide meaningful feedback on teaching quality for teachers in school practice.

This paper presents the application of classroom observations within a practical context. We conducted a systematic investigation of a teacher training that incorporated an observation system designed to provide formative feedback on teaching quality in school practice. Throughout the training, we tracked the quality of observer ratings of teaching quality at multiple time points—before, during, and after the training—and formulated validity arguments to support the use of these ratings as feedback on teaching quality. Specifically, we assessed the alignment with expert ratings, analyzed the factor structure of the ratings, and examined convergent correlations with a comparable observation system. Our study thereby demonstrates and evaluates an approach for training in-service teachers to generate valid observer ratings of teaching quality in school practice.

### Classroom Observations Using Standardized Observation Systems

Classroom observation by external observers is one of the main data sources for teaching-quality assessments. To systemize teaching-quality ratings resulting from classroom observations, there is increasing interest in using standardized classroom observation systems (Bell et al., 2019; Gitomer, 2021; Klette & Blikstad-Balas, 2018; Martinez et al., 2016; Praetorius & Charalambous, 2018). Standardized classroom observation systems are not only

protocols observers take notes on in classrooms but integrated systems for assessing distinct dimensions of teaching (Archer et al., 2016). These systems typically combine observation protocols with various other components (e.g., detailed observation manuals with scoring guidelines, rater trainings, and certifications; Bell et al., 2019; Hill et al., 2012). These different components serve to increase the quality of the resulting ratings (Bell et al., 2019).

Standardized classroom observation systems have been used for various purposes. In educational research, ratings of classroom videos using standardized classroom observation systems have deepened the understanding of teaching quality (Klieme et al., 2009; Pianta & Hamre, 2009) and its significance for student learning (e.g., Lipowsky et al., 2009; Lynch et al., 2017). In practice, standardized classroom observation systems have been used for teaching evaluations (Gitomer et al., 2021; Taut & Rakoczy, 2016) but also to improve teaching by providing formative feedback to teachers (Kraft & Hill, 2020; Muijs et al., 2018). Popular examples of classroom observation systems are different versions of the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008), the International Comparative Analysis of Learning and Teaching (ICALT; van de Grift, 2007), and the Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2013), which has been adapted for use in different subjects in Nordic countries (e.g., Stovner et al., 2021). Bell et al. (2019), Klette and Blikstad-Balas (2018), and Martinez et al. (2016) provide overviews of common classroom observation systems.

**Rater Training in Classroom Observation**

To use observation systems in a research context, raters usually need to undergo rater training (Bell et al., 2019; Martinez et al., 2016). In such trainings, observers get to know the system in detail and practice how to apply scoring rules to rate the quality of teaching, often by using classroom videos. Therefore, rater trainings are important for ensuring the quality of the ratings (Bell et al., 2019). For example, to use CLASS (Pianta et al., 2008), observers undergo a 2-day training by a certified CLASS trainer, followed by a reliability test and regular reliability checks through video coding, so-called calibration. CLASS observers need to recertify annually and receive additional support (e.g., webinars and individual mentoring; Head Start, 2023). To use PLATO (Grossman et al., 2013), there is a 5-day rater training (face-to-face or online) with a reliability test, where raters need to achieve 80% agreement with expert ratings in order to be certified to use the observation system. To use ICALT in different countries, raters undergo a rater training with three phases: preparation, implementation, and

evaluation (Maulana et al., 2021). To be certified, observers need to achieve 70% agreement both within the group and between the group and an expert rating (Maulana et al., 2021).

Although rater training is a well-established practice in research contexts, observation training for teachers in school practice remains less common and is scarcely studied (Gitzi et al., 2023; Martinez et al., 2016). When teachers use observation systems for feedback, there is no clear consensus on the amount of training needed to produce valid ratings of teaching quality. To our knowledge, the only systematic investigation of a rater training in classroom observation for practitioners is the study by Bergin et al. (2017), which assessed whether principals could accurately evaluate teaching quality after completing an observation training. While the study found that principals generally achieved high rating accuracy, their assessments varied depending on the specific teaching episodes and practices observed. Extending this line of research to teachers would be highly beneficial, as it remains unclear how their observer ratings develop during rater training. Understanding when and how rater training enhances the reliability and validity of these ratings would enable the design of more effective teacher training programs, ensuring that teachers provide valuable feedback on teaching quality. To address this, we systematically investigated a rater training in classroom observation for in-service teachers. In this training, teachers were trained to use an observation system aimed at developing teaching quality in German schools (Fauth et al., 2021).

**Validity of Teachers' Observer Ratings of Teaching Quality**

In this study, we examine the development of teachers' observer ratings of teaching quality throughout a training. Our primary goal is to determine whether these trained teachers can provide valuable feedback on teaching quality, based on the validity of their observer ratings. Therefore, we took a closer look at what the validity of observer ratings of teaching-quality means and how research could derive such validity arguments.

For a long time, validity was considered a property of a test (see Campbell & Fiske, 1959), but it is now defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association et al., 2014, p. 11). The process of validation is therefore considered the generation of evidence (validity arguments) for the legitimacy of test-score interpretation (Kane, 1992, 2013; Messick, 1995). Regarding standardized classroom observation systems, validity arguments need to support the intended use of the observer ratings, which means that different validity arguments may be appropriate for different uses of observations (e.g., Bell et al., 2012; Praetorius & Charalambous, 2018). For instance, if observer ratings are intended to provide

formative feedback on teaching quality, they must accurately reflect the quality of a specific aspect of teaching that teachers can understand and interpret effectively.

In looking at how studies on standardized classroom observation systems validated observer ratings, there is a very heterogeneous pattern that reflects the complexity of the validation process. In special education, for example, Rodgers et al. (2022) found that many studies reported no validity evidence for ratings at all. Studies addressing the validity of observer ratings of teaching quality show different approaches to validation. In their argument approach to observation protocol validity, Bell et al. (2012) emphasize that different strategies can be employed to construct a validity argument for observer ratings of teaching quality, depending on their intended interpretation and use. At the rater level, it is common to evaluate the consistency and accuracy of teaching-quality ratings as indicators of reliability and validity, such as measures of interrater agreement or alignment with expert ratings (e.g., Johnson et al., 2020). To measure construct validity, studies examined whether the factor structure of ratings aligns with the underlying theoretical framework of teaching (e.g., Li et al., 2020). As a measure of criterion validity, studies have tested associations with relevant outcomes related to teaching quality (e.g., Lynch et al., 2017). Another well-established approach to validation in psychological research involves establishing convergent validity with other measurement instruments that measure the same construct (e.g., Clark & Watson, 2019).

In summary, there are well-established ways to generate validity arguments for the interpretation and use of observer ratings of teaching quality: agreement with expert ratings, factor structure of the ratings, and associations with rating-relevant criteria (relevant outside criteria or ratings from comparable instruments). As researchers have called for clear statements of how studies derive validity arguments for teaching-quality ratings from standardized classroom observation systems (e.g., White, 2022), in this study, we aimed to formulate arguments for the validity of the observer ratings teachers gave over the course of a rater training. These ratings are intended to provide formative feedback on teaching quality to support professional development. Consequently, our validation process reflects on this intended use.

**Development of a Standardized Classroom Observation System for School Practice**

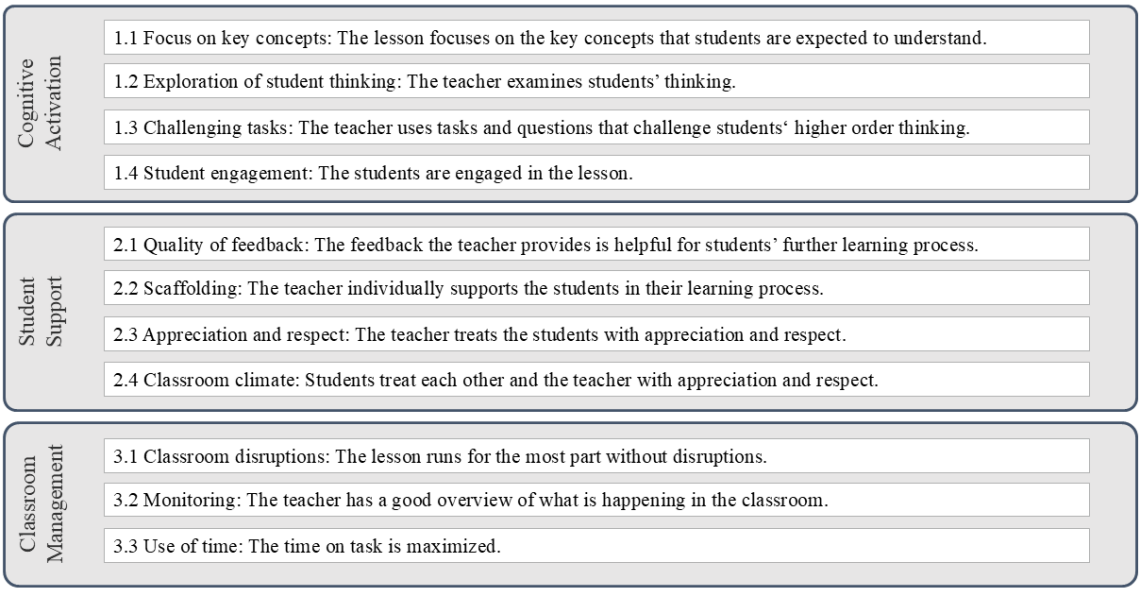
Observation systems used for professional development in school practice frequently lack a formal certification procedure, and there is often an absence of clear validity arguments supporting their use, particularly in German-speaking countries (Tarkian et al., 2019). To provide a system for classroom observation in school practice within German-speaking

countries, Fauth et al. (2021) developed a parsimonious, evidence-based observation system accompanied by a comprehensive teacher training.

The developed observation system is based on the framework of the three basic dimensions of teaching quality (Klieme et al., 2009; Praetorius et al., 2018): cognitive activation, student support, and classroom management. The three basic dimensions of teaching quality are a suitable theoretical framework for a feedback tool for practice, as these dimensions are very comprehensible and well-known by teachers in practice (e.g., Ruth-Herbein et al., 2022). In developing the observation system, Fauth et al. (2021) selected from each basic dimension the teaching-quality aspects that are both representative of the dimension and predictive of student learning (e.g., Fauth et al., 2014; Lipowsky et al., 2009). The resulting observation system consists of 11 items. As teaching quality is co-constructed by teachers and students (Fauth et al., 2020; Göllner et al., 2021), eight items refer to teachers' behavior, three items refer to students' behavior in the classroom, and each item corresponds to one basic dimension. The three items focusing on the students' behavior could be considered a separate method factor in the theoretical model of this system (Campbell & O'Connell, 1967). Figure 1 illustrates the 11 teaching-quality items and their assignment to the basic dimensions of teaching quality.

**Figure 1**

*Eleven Items From the Observation Form Assigned to the Three Basic Dimensions of Teaching Quality*



The observation system consists of the actual observation form with the 11 items and an accompanying observation manual. The manual is essential for the use of observation form, as it provides the theoretical background for the basic dimensions of teaching quality and items, as well as observable indicators. Furthermore, a rater training was developed to train teachers in classroom observation and provide feedback on teaching quality using the observation system. This training provides the context for the present investigation of the psychometric quality of teaching-quality ratings in rater trainings.

**Present Study**

In this study, we investigated observer ratings of teaching quality over the course of a classroom observation training aimed at enabling in-service teachers to effectively use a classroom observation system in school practice. Research Questions (RQs) 1 and 2 focused on examining the development of the reliability of teachers' ratings throughout and after the rater training. We asked:

**RQ1:** How does the reliability of teachers' observer ratings of teaching quality develop over the course of a rater training in classroom observation?

**RQ2:** Is the reliability of teachers' observer ratings of teaching quality satisfactory after the rater training?

In RQ3, we aimed to establish validity arguments supporting the use of these observer ratings by trained teachers as feedback on teaching quality in school practice (e.g., White, 2022). Through three subquestions (RQ3a, RQ3b, and RQ3c), we explored three different approaches to validating the observer ratings of teaching quality. We asked:

**RQ3**: Can we derive validity arguments for the use of teachers' observer ratings of teaching quality over the course of a rater training in classroom observation?

**RQ3a**: How does the agreement of teachers' ratings with expert ratings develop over the course of the rater training?

**RQ3b**: What is the factor structure of the teaching-quality ratings after the rater training?

**RQ3c**: Are the teaching-quality ratings associated with ratings from another classroom observation system for the same lessons?

## Method

This study is part of the project *Promoting Teaching Quality Through Classroom Observation and Feedback* of the Institute for Educational Analysis Baden-Württemberg and the Centre for School Quality and Teacher Education Baden-Württemberg.

### Participants

The present investigation was conducted with $N = 10$ experienced in-service teachers (50% women) who participated in the prepilot of the rater training for the classroom observation system by Fauth et al. (2021). The teachers worked as teacher educators in mathematics teacher education in the German state of Baden-Württemberg. At the time of the investigation, participants' professional experience ranged from 5 to 33 years ($M = 21.30$, $SD = 7.20$). Their teaching load in school practice ranged from 2 to 27 teaching hours per week ($M = 19.50$, $SD = 9.14$). Participants' experience in professional development for in-service teachers ranged from 2.5 to 21 years ($M = 11.45$, $SD = 6.84$), and their experience in pre-service teacher education ranged from 0 to 14 years (M = 7.40, SD = 5.78). The exact age of the participating teachers was not recorded for reasons of anonymity in the small sample. The teachers gave their informed consent to participate in the study.

### Rater Training Procedure

#### *Course of the Rater Training*

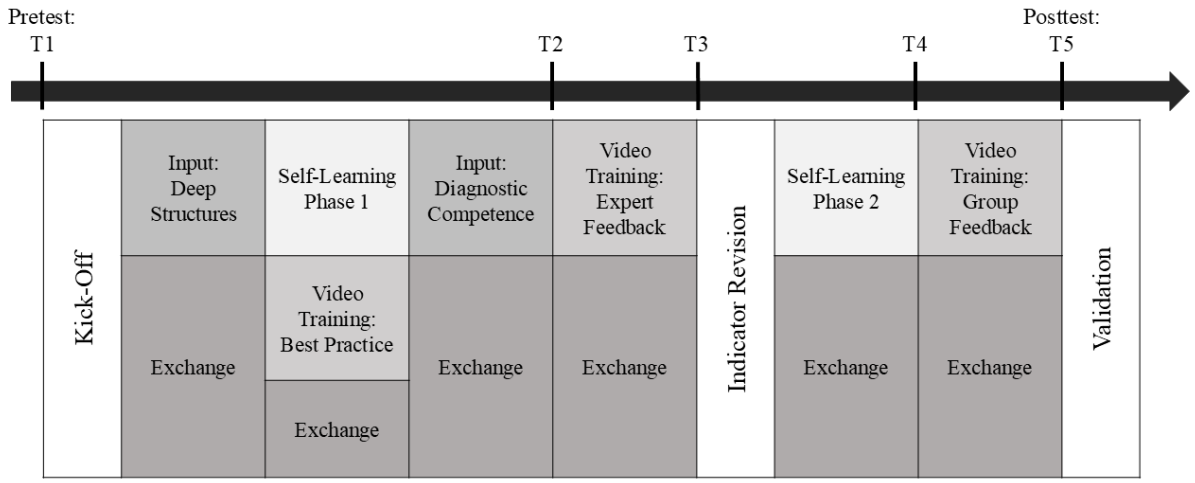
The rater training covered a total of 30 hr and was conducted by teaching-quality researchers from September to November 2020. Due to the COVID-19 pandemic, the rater training was carried out entirely online through the meeting tool BigBlueButton and the learning platform Moodle. Figure 2 provides a graphical overview of the rater training.

The training consisted of different synchronous and asynchronous training phases. In two theoretical input phases, participants received relevant theoretical background information on the topics *basic dimensions of teaching quality* and *diagnostic competence* (90 min each). In two self-learning phases (90 min and 60 min), participants worked independently with the observation form and the observation manual to get familiar with the system. In three video-training phases (one 30 min, two 90 min), participants observed short classroom video sequences and discussed ratings for the videos with the researchers conducting the training. Additionally, there were continuous exchange phases between the participating teachers and the trainers over the course of the rater training. In the exchange phases, the 11 teaching-quality items and positive and negative indicators in the observation manual were discussed using

concrete teaching examples. The aim of these exchange phases was to strengthen the common understanding of the items. Based on the group discussions, the indicators in the observation manual were revised once (indicator revision), and subject-specific observation guidelines were developed.

**Figure 2**

*Overview of the Rater Training and Time Points for the Teaching-Quality Ratings (T1–T5)*




In order to investigate the psychometric quality of the ratings, data were collected at five time points over the course of the training (T1–T5; see Figure 2). At these time points, participants independently rated short classroom video sequences (about five minutes each) using the observation system by Fauth et al. (2021). The first time point was the pretest (T1) that teachers took before starting the training. At T1, participants rated the quality of five classroom video sequences. T2, T3, and T4 consisted of ratings over the course of the training. Here, participants independently rated three (T2), 15 (T3), and five (T4) video sequences. After completing the rater training, the teachers independently rated 10 additional video sequences in a posttest (T5). All in all, participants rated a total of 38 classroom video sequences over the course of the rater training (T1–T5). At all five time points, videos were presented to the participants in a randomly assigned order. Teaching quality was rated on a 4-point Likert scale ranging from 1 (*not true*) to 4 (*totally true*). However, participants could choose an additional response category (5 = *not observed*) when the respective teaching-quality aspect was not observable in the video sequence.

The short classroom video sequences the participants rated over the course of the rater training were 4- to 5-min excerpts from classroom videos recorded in the TALIS video study

Germany (Grünkorn et al., 2020) in German secondary school classrooms. The videos show mathematics lessons on quadratic equations. Expert ratings of teaching quality with the observation system by Fauth et al. (2021) served as comparison data for the validity of the video ratings by the participating teachers. Expert ratings were available for eight of the 38 classroom video sequences (three videos from T2 and five videos from T4). These expert ratings were commonly created by two teaching-quality researchers.

### *Validation Study After the Rater Training*

After the rater training was completed, the 10 teachers participated in a follow-up validation study from mid-November 2020 to early February 2021. During this period, the teachers observed 45-min classroom videos from the Pythagoras study (Klieme et al., 2009). These videos show 45-min mathematics lessons on the Pythagorean Theorem. Participants rated the teaching quality in the videos using the observation form they were trained to use. The total number of 37 Pythagoras videos was divided into six rating blocks. The first rating block with three classroom videos served as the training block for teachers to get familiar with rating longer periods of teaching (45-min classroom videos) and was not analyzed. The final sample of 34 classroom videos rated in the validation study consisted of the videos rated in Rating Blocks 2–6. All participants rated all videos from Rating Blocks 1–5 (22 videos). Due to the high time demands for participating in the study in the last rating block, the teachers rated only five of the 12 videos. The videos were evenly distributed and randomly assigned to the teachers, so that each video in Rating Block 6 was rated by four to five teachers.

Over the course of the validation study, teachers participated in both synchronous and asynchronous phases of exchange with each other and with the experts who conducted the rater training before. At five synchronous online meetings, participants could reflect on the previously rated classroom videos and get clarification on uncertainties regarding the teaching-quality ratings with the observation system. Additionally, the online platform Crypt-Pad offered an asynchronous option for exchanging rating experiences.

In the validation study, the original teaching-quality ratings from the Pythagoras study (Klieme et al., 2009) served as comparison data for the validity of the video ratings by the participating teachers. The ratings from the Pythagoras study have served as valid indicators of teaching quality in previous studies (e.g., Lipowsky et al., 2009). For this reason, we assume that these ratings provide appropriate comparison data for the validity of the observer ratings of teaching quality in our study. To use the teaching-quality ratings from the Pythagoras study as comparison data, we conducted a theoretical comparison of the observation systems used in

the Pythagoras study (Rakoczy & Pauli, 2006) and this study (Fauth et al., 2021) by assigning the items that assessed the same or very similar aspects of teaching quality. After the 10 teachers participating in the validation study rated the 34 classroom videos from the Pythagoras study, the teaching-quality ratings for all of the videos were available from both classroom observation systems.

**Statistical Analyses**

Data analyses were conducted with R version 4.4.0 and IBM SPSS Statistics version 27. To address RQ1 and RQ2, we investigated the reliability of teaching-quality ratings over the course of a teacher training in classroom observation. To measure reliability, we used interrater agreement between raters (LeBreton & Senter, 2008). We estimated interrater agreement with the Average Absolute Deviation Index ($AD_M$), indicating the average absolute deviation of the raters from the group mean (Burke et al., 1999). The smaller the $AD_M$ value is, the higher the agreement between raters. We used the $AD_M$ because it is sensitive to rating variability while also offering a user-friendly interpretation, making it well-suited for practical contexts such as teacher training (Burke et al., 1999). To investigate how reliability developed over the course of the rater training (RQ1), we examined the $AD_M$ values over the course of the rater training (T1–T5). To determine whether the reliability of teaching-quality ratings was satisfactory after the rater training (RQ2), we considered whether the $AD_M$ values at posttest (T5) reached a predefined cut-off criterion. In line with Smith-Crowe et al. (2014), we used a cut-off value of 0.60. We excluded cases from the reliability analysis when more than three participants indicated that they did not observe the respective teaching-quality aspect in the video sequence (additional response category 5).

To investigate the validity of the ratings in RQ3, we used different analytic approaches. To examine the agreement between the teachers' ratings and expert ratings (RQ3a), we used an adjusted $AD_M$, indicating the average absolute deviation of the teachers' ratings from the expert ratings and considered the development between T2 and T4.

To examine the factor structure of the ratings (RQ3b), we conducted a multilevel confirmatory factor analysis (MCFA) for three theory-based possible factor models: one global factor model with all 11 teaching-quality items forming a global teaching-quality factor (Model 1), one model with the 11 items forming the three basic dimensions of teaching quality (Model 2), and one model with the 11 items forming the three basic dimensions of teaching quality and an additional uncorrelated factor consisting of the three student-focused items as a method factor (Model 3). As raters were nested in classroom videos, the videos served as a cluster

variable in the multilevel structure. To evaluate the goodness of fit for the three suggested models, we used the criteria by Hu and Bentler (2009), who suggested that the Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) should be close to .95 and the Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) should be close to .08.[1] After evaluating the model fits of the three suggested models, we compared the different models using $\chi^2$ difference tests, the Akaike Information Criterion (AIC), and the Bayes Information Criterion (BIC), where models with lower values are preferred (Raftery, 1993).

To examine associations with ratings from another classroom observation system for the same teaching lessons (RQ3c), we examined correlations between the items assessing the same or very similar teaching-quality aspects from the observation systems from the Pythagoras study and this study for the classroom videos from the Pythagoras study (Klieme et al., 2009). In addition to the correlations on the item level, we investigated the correlations between the observation systems on the scale level, aggregating the items into preassigned scales. We compared the correlations between scales within and between the two observation systems using a multitrait-multimethod approach (MTMM; Campbell & Fiske, 1959). The MTMM analysis allowed us to systematically examine convergent and discriminant correlations between the teaching-quality dimensions from the two instruments. As an indicator of validity, we expected convergent correlations to be higher than discriminant correlations (Campbell & Fiske, 1959; Schmitt & Stults, 1986). The alpha level was set at .05 in accordance with common practice.

[1] When comparing the results of the multilevel confirmatory factor analysis, the level-specific SRMR was used to obtain information about the model fit at both the within and between levels. Due to the small number of videos we used in the present study, it can be assumed that the $SRMR_{between}$ was not sensitive enough to detect potential misspecification. For this reason, we used traditional cut-off values to evaluate within-level models (Lin & Hsu, 2022). Additionally, we performed a $\chi^2$ difference test to make a clear decision between the competing models.

## Results

### Descriptive Results

Table 1 presents descriptive statistics for the 11 teaching-quality items from the observation system for the 38 classroom video sequences in the rater training. The high mean values indicate that teaching quality was rated high in the short video sequences, especially for the items *focus on key concepts*, *appreciation and respect*, *classroom climate*, *classroom disruptions*, *monitoring*, and *use of time* with mean values above 3.

**Table 1**

*Descriptive Statistics for the Teaching-Quality Ratings for the 38 Classroom Video Sequences in the Rater Training*

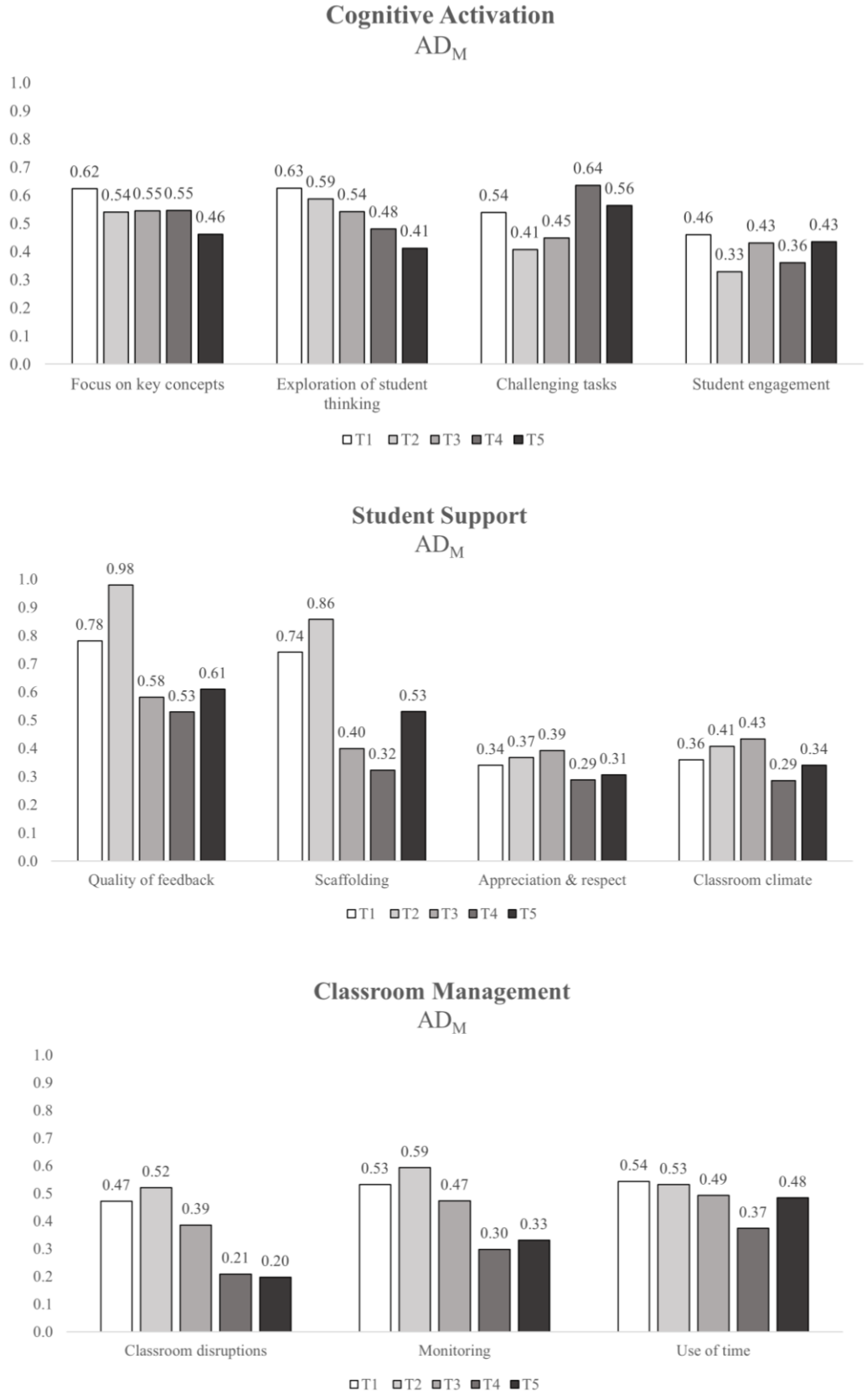
| Teaching-quality aspects | *M* | *SD* | Min | Max |
|---|---|---|---|---|
| Cognitive activation | | | | |
| Focus on key concepts | 3.30 | 0.86 | 1 | 4 |
| Exploration of student thinking | 2.76 | 1.07 | 1 | 4 |
| Challenging tasks | 2.70 | 1.03 | 1 | 4 |
| Student engagement | 2.91 | 0.89 | 1 | 4 |
| Student support | | | | |
| Quality of feedback | 2.69 | 0.98 | 1 | 4 |
| Scaffolding | 2.74 | 1.05 | 1 | 4 |
| Appreciation and respect | 3.44 | 0.79 | 1 | 4 |
| Classroom climate | 3.40 | 0.83 | 1 | 4 |
| Classroom management | | | | |
| Classroom disruptions | 3.41 | 0.89 | 1 | 4 |
| Monitoring | 3.19 | 0.92 | 1 | 4 |
| Use of time | 3.17 | 0.88 | 1 | 4 |

### Rater Training

RQ1 asked how the reliability of teachers' observer ratings of teaching quality develops over the course of a rater training. Figure 3 shows the development of interrater agreement ($AD_M$) over the course of the five time points T1–T5 for the 11 teaching-quality items from the observation system.

**Figure 3**

*Interrater Agreement $AD_M$ Over the Course of the Rater Training (T1–T5)*



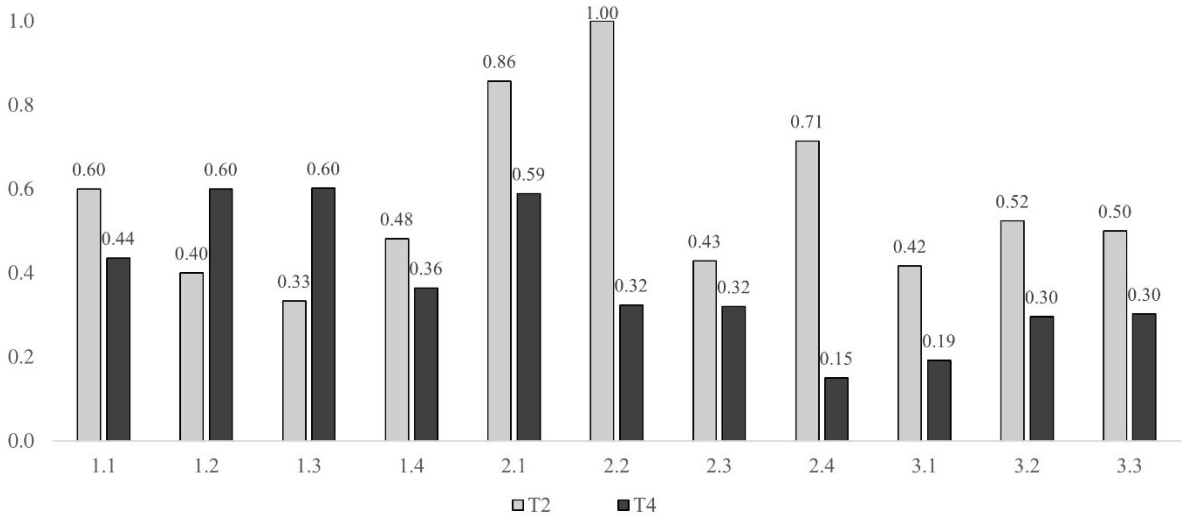
We found that the $AD_M$ values decreased, meaning that rater agreement improved as the training progressed. For the items *focus on key concepts* and *exploration of student thinking*, we found a linear increase in agreement. For most items, however, we found heterogeneous development in the interrater agreement with different patterns of increasing and decreasing agreement between raters at the different time points. It is particularly striking that the items *quality of feedback* and *scaffolding* from the student support dimension showed little agreement in ratings at the beginning of the training (T1 and T2), but agreement increased over the course of the training. In addition, for specific items, the raters already demonstrated high agreement from the beginning, especially for the items *student engagement*, *appreciation and respect*, *classroom climate*, and the three items from the classroom management dimension: *classroom disruptions*, *monitoring*, and *use of time*. Even though most items did not show a linear increase in agreement, there was a visible trend toward greater agreement over the course of the training.

RQ2 asked if the reliability of teaching-quality ratings was satisfactory after the rater training. We compared the $AD_M$ values at posttest (T5) to the predefined value of 0.60 or below as the cut-off criterion for satisfactory interrater agreement. We found that the raters achieved satisfactory interrater agreement at posttest for all teaching-quality items, except for the item *quality of feedback* ($AD_M = 0.61$; see Figure 3).

RQ3's three subquestions asked for validity arguments for the use of teaching-quality ratings of the trained teachers. RQ3a asked how the agreement between the teachers' ratings and the expert ratings developed over the course of the training. We found that the agreement with expert ratings increased for nine of the 11 teaching-quality items between T2 and T4 (see Figure 4). For two of the items, *exploration of student thinking* and *challenging tasks*, agreement with the expert ratings decreased. It is particularly striking that the items *quality of feedback* and *scaffolding* again showed little agreement at T2 but, remarkably, agreement with experts was higher at T4.

**Figure 4**

*Average Deviations From Expert Ratings at T2 and T4*




*Note.* 1.1 = Focus on key concepts, 1.2 = Exploration of student thinking, 1.3 = Challenging tasks, 1.4 = Student engagement, 2.1 = Quality of feedback, 2.2 = Scaffolding, 2.3 = Appreciation and respect, 2.4 = Classroom climate, 3.1 = Classroom disruptions, 3.2 = Monitoring, 3.3 = Use of time.

## Validation Study

RQ3b asked about the factor structure of teaching-quality ratings after the rater training. We tested three theory-based possible factor models for our data using multilevel CFA: a global factor (Model 1), a model with three basic dimensions (Model 2), and a model with three basic dimensions and an additional method factor consisting of the student-focused items (Model 3). Table 2 presents the fit indices of the multilevel CFA for the three proposed factor models. The fit indices indicated that Model 3 fit the data best. Model 3's fit indices supported a relatively good model fit according to the criteria suggested by Hu and Bentler (2009), with the exception of the between-level SRMR value (.157), which exceeded the suggested value of .08 (Hu & Bentler, 2009), as expected. Comparing the models using $\chi^2$ difference tests, we found that Model 3 fit the data significantly better than Model 2 ($p < .001$) or Model 1 ($p < .001$), and the AIC and BIC values agreed. Thus, the ratings after the rater training appeared to represent the three basic dimensions of teaching quality, with the three student-focused items (*student engagement*, *classroom climate*, and *classroom disruptions*) representing an additional separate method factor.

**Table 2**

*Fit Indices From the Multilevel CFA for the Three Factor Models*

| | M1<br>Global Factor Model | M2<br>Model With Three Basic Dimensions | M3<br>Model With Three Basic Dimensions and a Separate Method Factor (Student-Focused Items) |
|---|---|---|---|
| $\chi^2(df)$ | 189.942 (88) | 146.062 (82) | 115.039 (80) |
| CFI | .871 | .919 | .956 |
| TLI | .839 | .891 | .939 |
| RMSEA | .080 | .066 | .049 |
| $SRMR_W$ | .096 | .088 | .068 |
| $SRMR_B$ | .212 | .166 | .157 |
| AIC | 3464.512 | 3432.632 | 3405.61 |
| BIC | 3640.430 | 3627.74 | 3607.115 |
| ABIC | 3466.241 | 3434.549 | 3407.590 |

*Note.* Level 1 = Rater; Level 2 = Videos; $_W$ = within (Level 1); $_B$ = between (Level 2).

RQ3c asked whether the participants' ratings of the classroom videos from the Pythagoras study were associated with the original ratings from the Pythagoras study. After assigning items that assessed the same or similar teaching-quality aspects, we examined the convergent correlations between the preassigned items from the two systems for the rated classroom videos from the Pythagoras Study (see Table 3). We found moderate to large convergent correlations between the preassigned teaching-quality items, with the exception of the item *focus on key concepts* (Fauth et al., 2021), the two assigned items *learning status awareness* and *clarity* (Rakoczy & Pauli, 2006), and *exploration of student thinking* (Fauth et al., 2021) and the assigned item *exploration of prior knowledge* (Rakoczy & Pauli, 2006), for which there were no significant correlations. The significant convergent correlations ranged from $r = .30–.78$.

**Table 3**

*Convergent Correlations Between the Preassigned Items From the Two Observation Systems for the Classroom Videos From the Pythagoras Study*

| Ratings in Validation Study (Fauth et al., 2021) | Original Ratings in Pythagoras Study (Rakoczy & Pauli, 2006) | Correlation |
|---|---|---|
| Cognitive activation | | |
| Focus on key concepts | Learning status awareness | .09 |
| | Clarity | .08 |
| Exploration of student thinking | Exploration of prior knowledge | .22 |
| | Exploration of student thinking | .63** |
| Challenging tasks | Challenging problems | .30* |
| Student engagement | Student engagement | .47** |
| Student support | | |
| Quality of feedback | Factual-constructive feedback | .55** |
| | Positive culture of error | .40* |
| Scaffolding | Individualization | .40* |
| Appreciation and respect | Recognition of the teacher | .55** |
| Classroom climate | Learning community | .32* |
| | Recognition of the students | .46** |
| Classroom management | | |
| Classroom disruptions | Discipline problems | .78** |
| Monitoring | Classroom management | .55** |
| Use of time | Classroom management | .45** |

* $p < .05$. ** $p < .01$.

We also considered the associations between the preassigned items aggregated to scales, in order to compare convergent and discriminant correlations systematically using an MTMM approach. Table 4 presents the MTMM matrix summarizing the correlations within and between the two observation systems by Fauth et al. (2021) and Rakoczy and Pauli (2006). The blocks indicating the correlations of the scales within an observation system are shaded in gray. The reliability diagonal, which reports Cronbach's alpha for the respective scale as a measure of internal consistency, is shaded in dark gray. For the scales from Fauth et al.'s (2021) system, these were satisfactory (α = .82–.87). For the scales formed based on the preassigned items by Rakoczy and Pauli (2006), the values ranged from .43 to .99. With α = .43, the cognitive activation scale was not sufficiently reliable. By contrast, the reliability values for student support and classroom management were satisfactory. The values in bold represent the convergent correlations between the scales from the two observation systems. The convergent correlations ranged from .55 to .67. With values > .50, these correlations were consistently strong according to Cohen (1988). Although significant discriminant correlations were evident, the convergent correlations were consistently higher.

**Table 4**

*Multitrait-Multimethod Matrix on the Scale Level*

| | | Fauth et al. (2021) | | | Rakoczy and Pauli (2006) | | |
|---|---|---|---|---|---|---|---|
| | | CA | SS | CM | CA | SS | CM |
| Fauth et al. (2021) | CA | .87 | | | | | |
| | SS | .74** | .82 | | | | |
| | CM | .44* | .64** | .83 | | | |
| Rakoczy and Pauli (2006) | CA | **.55**** | .40* | .17 | .43 | | |
| | SS | .51** | **.67**** | .22 | .44* | .81 | |
| | CM | .40* | .57** | **.63**** | .36* | .41* | .99 |

*Note.* CA = Cognitive Activation, SS = Student Support, CM = Classroom Management. Cronbach's alpha reliability values are presented on the diagonal. Convergent correlations are in bold.

* $p < .05$. ** $p < .01$.

## Discussion

The present study conducted a systematic investigation of a teacher training in classroom observations. We examined the development of teachers' observer ratings during this training and presented three validity arguments supporting the use of these observer ratings by the trained teachers as feedback on teaching quality in school practice (White, 2022).

### Summary and Interpretation of Findings

In the first part of this study, we monitored the quality of teaching-quality ratings over the course of a rater training. We found that overall interrater agreement between the participating teachers increased over the course of the training (RQ1). However, there were item-specific differences in how the ratings developed, with some showing linear improvements in agreement, others fluctuating, and a few maintaining consistently high levels of agreement throughout the training. A possible explanation for the observed developments is that the discussions in the exchange phases of the rater training influenced how the raters understood and rated the teaching-quality items. This suggestion is supported by the fact that this mixed pattern of decreases and increases in agreement was especially evident for the items from the cognitive activation dimension and the items *quality of feedback* and *scaffolding* from the student support dimension, which were extensively discussed in the exchange phases. This finding about different rating quality for different teaching-quality aspects is in alignment with other studies that pointed out that specific items, mostly related to instructional dimensions of observation protocols, are especially challenging to rate consistently (Bell et al., 2015; Bergin et al., 2017; Kane & Staiger, 2012; Maulana et al., 2021). From a theoretical perspective, it is plausible that different aspects of teaching quality vary in how well they can be assessed through observation (Fauth et al., 2020). For example, the assumptions underlying the cognitive activation dimension are based on cognitive constructivist learning theories (e.g., Piaget, 1955) and focus on a high level of cognitive processing of learning content. This is less observable from the outside than, for instance, aspects of classroom management, where time on task (Carroll, 1963) is central and can be more readily observed.

With RQ2, we asked if agreement at posttest (T5) was satisfactory and found satisfactory agreement after the rater training for all items, except for *quality of feedback*. This finding confirms that the rater training succeeded in creating a common understanding of teaching quality and its assessment, which is one of the aims of the use of standardized classroom observation systems (Praetorius & Charalambous, 2018). However, the item *quality of feedback*, as one of the most discussed items in the rater training, needs further training for

teachers to achieve the standards of satisfactory consistency. Rater trainings need to address this issue and include more elements to address the meaning of the aspect *quality of feedback* in trainings.

All in all, the development of interrater agreement between teachers revealed that the rater training with its different learning and exchange phases was successful in creating a common understanding of teaching quality, but for specific items the ratings were more challenging than for others (e.g., Bergin et al., 2017). Based on our results, we suggest special attention in rater trainings for aspects related to cognitive activation and the student-support-related items that are more focused on the cognitive facet of student support (e.g., *quality of feedback* and *scaffolding*; Kleickmann et al., 2020). In comparison, more affective facets of student support (e.g., *appreciation and respect* or *classroom climate*) and the aspects of classroom management were rated with high agreement from the beginning. Additionally, it is important to note that achieving high agreement among all raters should not be the sole objective of training teachers to provide observer ratings. Instead, the focus should be on fostering a common language of teaching quality (Klette, 2023) and ensuring that ratings serve as useful information for enhancing teaching quality in school practice (Bell & Gitomer, 2023).

We generated three validity arguments for the use of teaching-quality ratings of the trained teachers in RQ3's subquestions. In investigating the agreement with expert ratings, we found that the agreement increased with time for nine of the 11 items (RQ3a). This finding indicates that the experts' and the teachers' understanding of the teaching-quality ratings became more similar over the course of the rater training, except for the two items that seemed to need further training. The increasing agreement with expert ratings can be used as a first argument for the validity of the observer ratings of teaching quality as useful feedback on teaching quality in school practice (Cook et al., 2015; Johnson et al., 2020).

In the second part of this study, we investigated the ratings of trained teachers in a validation study after the rater training. In investigating three possible factor models (RQ3b), we found that a model with the 11 items representing the three basic dimensions of teaching quality and an additional, uncorrelated method factor consisting of the three student-focused items fit the data best. We conclude that the different focus of observation for student-focused items in observation systems is not negligible and needs to be considered as a method factor to have a valid interpretation of the ratings (Campbell & O'Connell, 1967). For the conceptualization of teaching quality, this result makes clear that student-focused aspects of teaching quality, such as student engagement (e.g., Fredricks et al., 2019), classroom climate

(e.g., Schweig et al., 2019), or disruptive behavior (e.g., Scherzinger & Wettstein, 2019), are substantial for the conceptualization of teaching quality (Fauth et al., 2020; Göllner et al., 2021). The factor analysis of the ratings in the validation study can be used as a second validity argument for the use of the teachers' ratings (Li et al., 2020), as the different teacher- and student-focused items to assess teaching quality build useful feedback for central teacher- and student-focused teaching-quality aspects (Bell et al., 2012).

In investigating the associations with teaching-quality ratings for the same classroom videos with different observation systems (RQ3c), we found mostly moderate to high convergent correlations and, using a MTMM matrix, we found that convergent correlations were higher than discriminant correlations, supporting the validity of the teaching-quality ratings conducted by trained teachers in a rating phase after the training (Campbell & Fiske, 1959; Clark & Watson, 2019; Schmitt & Stults, 1986). The moderate to high convergent correlations between items with the same or similar content were remarkable considering that the ratings from the observation systems used in the Pythagoras study stemmed from a different rating context (Klieme et al., 2009).

**Limitations**

Our study has several limitations. First, there were different limitations in the study design and the rater-training procedure, which needed to fit the conditions of a training under practice conditions for in-service teachers with limited time resources. As this rater training was the prepilot trial of the training for a newly developed observation system, we had only a small sample of ten teachers. The small and specific sample of highly engaged in-service teachers may limit the generalizability of the results. Furthermore, the sample size was especially critical for the factor analyses we conducted (e.g., Hox & Maas, 2001). Whereas our findings indicated that a model with three teaching-quality dimensions and a separate method factor for student-focused items was superior to less faceted models, the sample size might not have been large enough to conduct sensitive tests, particularly for identifying the best-fitting models at the video level. Future research with a larger number of videos will be needed to conduct more robust tests to assess the appropriateness of a theoretically driven three-factor solution with a separate method factor.

As an additional limitation, the number of classroom videos rated at the different time points during rater training varied. For this reason, interrater agreement was based on different numbers of observations at the different time points. As comparison data for the videos used during rater training, we used expert assessments of the videos at T2 and T4 due to their

availability. However, it would have been interesting to examine the development of agreement with expert ratings from pretest (T1) to posttest (T5).

Another factor to consider are the standards we used to determine a measure's validity. We used expert ratings and the original Pythagoras ratings (Klieme et al., 2009) as comparison standards for our data. However, we need to critically consider whether these measures really provide a true teaching-quality score and whether such a true score even exists (Fauth et al., 2020). Regarding the original ratings from the Pythagoras study, it is necessary to recognize that the ratings used for this comparison came from different contexts. The ratings from the Pythagoras study stemmed from a different group, a different number of raters, and a different time (almost 20 years ago). These differences might have influenced the ratings, as the meaning of some teaching-quality aspects may have changed over the last 20 years (Gossner et al., 2023).

**Significance and Implications**

In this study, we investigated a rater training in classroom observation using an observation system for teaching-quality development in school practice. This research contributes to a much-needed deeper understanding of teacher trainings in classroom observation (Gitzi et al., 2023). Moreover, we provide an example of how to derive validity arguments for teachers' ratings of teaching quality in school practice (White, 2022) and introduced an observation system suitable as a tool for providing feedback to teachers in school practice accompanied by a teacher training.

The results and implications of this study are especially valuable for researchers and practitioners training observers in classroom observation systems (e.g., Head Start, 2023; Maulana et al., 2021). One important implication is that teacher trainings in classroom observation have the potential to foster a shared understanding of how to rate teaching quality (Klette, 2023); however, a linear improvement in rating quality should not be anticipated. Other important implications are that there are clear item-specific differences in the development of the psychometric quality of ratings, and it is worth closely monitoring the quality of ratings over the course of the training to see which items or dimensions need special attention in the rater training (e.g., Park et al., 2014). In alignment with previous research, our results indicate that particularly teaching-quality aspects from the cognitive activation and student support dimensions need special attention in trainings. Moreover, when observation systems contain some items focusing on teacher behavior and some focusing on student behavior, the different focus of observation between the items may affect the ratings, and this difference should be considered when interpreting the ratings (White & Klette, 2023).

All in all, this investigation can serve as a starting point for more systematic research on teacher trainings in classroom observation and be used as an example of how to design, evaluate, and report rater trainings with a pre-posttest design. To better understand the effectiveness of rater trainings, future studies on rater trainings should use control-group designs where specific aspects of the rater training (e.g., duration, training phases) are manipulated to learn more about which training components are especially effective for improving teaching-quality ratings. Additionally, further studies in practical contexts should examine whether observer ratings provided by teachers as formative feedback result in an improved quality of teaching.

## References

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

Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations*. John Wiley & Sons.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Bell, C. A., & Gitomer, D. H. (2023). Building the field's knowledge of teaching and learning: Centering the socio-cultural contexts of observation systems to ensure valid score interpretation. *Studies in Educational Evaluation*, *78*, 101278. https://doi.org/10.1016/j.stueduc.2023.101278

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62–87. https://doi.org/10.1080/10627197.2012.715014

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Mccaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2015). Improving observational score quality: Challenges in observer thinking. Designing teacher evaluation systems: New guidance from the measures of effective teaching project. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (pp. 50–97). John Wiley & Sons. https://doi.org/10.1002/9781119210856.ch3

Bergin, C., Wind, S. A., Grajeda, S., & Tsai, C. L. (2017). Teacher evaluation: Are principals' classroom observations accurate at the conclusion of training?. *Studies in Educational Evaluation*, *55*, 19–26. https://doi.org/10.1016/j.stueduc.2017.05.002

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, *2*(1), 49–68. https://doi.org/10.1177/109442819921004

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., & Schmidt, W. (2019). A review of the literature on teacher effectiveness and student outcomes. In N. Burroughs, J. Gardner, Y. Lee, S. Guo, I. Touitou, K. Jansen, & W. Schmidt (Eds.), *Teaching for excellence and equity: Analyzing teacher characteristics, behaviors and student outcomes with TIMSS* (pp. 7–17). Springer. https://doi.org/10.1007/978-3-030-16151-4_2

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, D. T., & O'Connell, E. J. (1967). Methods factors in multitrait-multimethod matrices: Multiplicative rather than additive?. *Multivariate Behavioral Research*, *2*(4), 409–426. https://doi.org/10.1207/s15327906mbr0204_1

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*, 723–733.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hoboken: Taylor and Francis.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. https://doi.org/10.1111/medu.12678

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who Sees What?: Conceptual Considerations on the Measurement of Teaching Quality from Different Perspectives. *Zeitschrift für Pädagogik Beiheft*, *1*, 138–155. https://doi.org/10.3262/ZPB2001138

Fauth, B., Herbein, E., & Maier, J. L. (2021). Beobachtungsmanual zum Unterrichtsfeedbackbogen Tiefenstrukturen [Observation manual for the classroom feedback form deep structures]. Institut für Bildungsanalysen Baden-Württemberg.

Fredricks, J. A., Reschly, A. L., & Christenson, S. L. (2019). Interventions for student engagement: Overview and state of the field. In J. A. Fredricks, A. L. Reschly, & S. L. Christenson (Eds.), *Handbook of student engagement interventions: Working with disengaged students* (pp. 1–12). Academic Press.

Gitomer, D. H. (2021). Methods for Observing Classroom Interaction. In R. Coe, M. Waring, L. Hedges, & L. Day Ashley (Eds.), *Research Methods and Methodologies in Education, 3rd Edn* (pp. 221–231). SAGE Publications.

Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2021). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, *58*(1), 3–31. https://doi.org/10.3102/0002831219890608

Gitzi, V., Wemmer-Rogh, W., Kleickmann, T., Lichtner, O., Steffensky, M., Heinze, A., & Praetorius, A.-K. (2023). *Effektivität von Trainings für Rater*innen zur Beurteilung von Unterrichtsqualität: Ein Literaturüberblick [Effectiveness of Rater Trainings to Assess Teaching Quality: A Literature Review]* [Paper presentation]. 10. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Essen, Germany.

Göllner, R., Fauth, B., & Wagner, W. (2021). Student Ratings of Teaching Quality Dimensions: Empirical Findings and Future Directions. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers* (pp. 111–122). Springer International Publishing. https://doi.org/10.1007/978-3-030-75150-0_7

Gossner, L., Wemmer-Rogh, W., Schreyer, P., Grob, U., Klieme, E., & Praetorius, A.-K. (2023). *Teaching quality What has changed in the last twenty years?* [Paper presentation]. 20th Biennial EARLI Conference, Thessaloniki, Greece.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445–470. https://doi.org/10.1086/669901

Grünkorn, J., Klieme, E., Praetorius, A.-K., & Schreyer, P. (2020). *Mathematikunterricht im internationalen Vergleich. Ergebnisse aus der TALIS-Videostudie Deutschland. [Mathematics instruction in international comparison. Results from the TALIS video study Germany]*. Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF). https://doi.org/10.25656/01:21156

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

HEAD Start (2023). *Use of Classroom Assessment Scoring System (CLASS®) in Head Start.* Retrieved March 6, 2024 from https://eclkc.ohs.acf.hhs.gov/designation-renewal-system/article/use-classroom-assessment-scoring-system-class-head-start

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hu, L., & Bentler, P. M. (2009). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2020). Examining rater accuracy and consistency with a special education observation protocol. *Studies in Educational Evaluation*, *64*, 100827. https://doi.org/10.1016/j.stueduc.2019.100827

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

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

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Bill and Melinda Gates Foundation. Retrieved March 6, 2024 from https://files.eric.ed.gov/fulltext/ED540960.pdf

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. C. (2020). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, *28*(62), 1–34. https://doi.org/10.14507/epaa.28.5012

Kleickmann, T., Steffensky, M., & Praetorius, A. K. (2020). Quality of teaching in science education. More than Three Basic Dimensions? *Zeitschrift Für Pädagogik*, *66*, Beiheft, 37–55. https://doi.org/10.25656/01:25862

Klette, K. (2023). Classroom observation as a means of understanding teaching quality: towards a shared language of teaching?. *Journal of Curriculum Studies*, *55*(1), 49–62. https://doi.org/10.1080/00220272.2023.2172360

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal*, *17*(1), 129–146. https://doi.org/10.1177/1474904117703228

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik, & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Kraft, M. A., & Hill, H. C. (2020). Developing ambitious mathematics instruction through web-based coaching: A randomized field trial. *American Educational Research Journal*, *57*(6), 2378–2414. https://doi.org/10.3102/0002831220916840

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Li, H., Liu, J., & Hunter, C. V. (2020). A meta-analysis of the factor structure of the classroom assessment scoring system (CLASS). *The Journal of Experimental Education*, *88*(2), 265–287. https://doi.org/10.1080/00220973.2018.1551184

Lin, J. J., & Hsu, H. Y. (2022). Investigating the performance of level-specific fit indices in multilevel confirmatory factor analysis with dichotomous indicators: A Monte Carlo study. *Behavior Research Methods*, 1–38. https://doi.org/10.3758/s13428-022-02014-z

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537. https://doi.org/10.1016/j.learninstruc.2008.11.001

Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between observations of elementary mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education*, *123*(4), 615–646. https://doi.org/10.1086/692662

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation*, 49, 15–29. https://doi.org/10.1016/j.stueduc.2016.03.002

Maulana, R., André, S., Helms-Lorenz, M., Ko, J., Chun, S., Shahzad, A., Irnidayanti, Y., Lee, O., de Jager, T., Coetzee, T., & Fadhilah, N. (2021). Observed teaching behaviour in secondary education across six countries: measurement invariance and indication of cross-national variations. *School Effectiveness and School Improvement*, *32*(1), 64–95. https://doi.org/10.1080/09243453.2020.1777170

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Muijs, D., Reynolds, D., Sammons, P., Kyriakides, L., Creemers, B. P., & Teddlie, C. (2018). Assessing individual lessons using a generic teacher observation instrument: how useful is the International System for Teacher Observation and Feedback (ISTOF)?. *ZDM*, *50*, 395–406. https://doi.org/10.1007/s11858-018-0921-9

Panayiotou, A., Herbert, B., Sammons, P., & Kyriakides, L. (2021). Conceptualizing and exploring the quality of teaching using generic frameworks: A way forward. *Studies in Educational Evaluation*, *70*, 101028. https://doi.org/10.1016/j.stueduc.2021.101028

Park, Y. S., Chen, J., & Holtzman, S. L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation system: New guidance from the measures of effective teaching project* (pp. 383–414). San Francisco, CA: Jossey-Bass A Wiley Brand. https://doi.org/10.1002/9781119210856.ch12

Piaget, J. (1955). The development of time concepts in the child. In P. H. Hoch & J. Zubin (Eds.), *Psychopathology of childhood* (pp. 34–44). New York, NY: Grube and Stratton

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System.* Baltimore, MD: Brookes.

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of the three basic dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A. K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 163–180). Newbury Park: Sage.

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse [Highly-inferent rating: Assessing the quality of teaching processes]. In I. Hugener, E. Klieme, C. Pauli, & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". 3. Videoanalysen* (pp. 206–233). GFPF, Deutsches Institut für Internationale Pädagogische Forschung (DIPF). https://doi.org/10.25656/01:3130

Rodgers, W. J., Morris-Mathews, H., Romig, J. E., & Bettini, E. (2022). Observation studies in special education: A synthesis of validity evidence for observation systems. *Review of Educational Research*, *92*(1), 3–45. https://doi.org/10.3102/00346543211042419

Ruth-Herbein, E., Maier, J. L., Fauth, B. (2022). Promoting Teaching Quality Through Classroom Observation and Feedback: Design of a Program in the German State of Baden-Württemberg. In J. Manzi, Y. Sun, & M. R. García (Eds.) *Teacher Evaluation Around the World. Teacher Education, Learning Innovation and Accountability.* (pp. 271–289). Springer, Cham. https://doi.org/10.1007/978-3-031-13639-9_12

Schmitt, N., & Stults, D. M. (1986). Methodology Review: Analysis of Multitrait-Multimethod Matrices. *Applied Psychological Measurement*, *10*(1), 1–22. https://doi.org/10.1177/014662168601000101

Scherzinger, M., Wettstein, A. (2019). Classroom disruptions, the teacher-student relationship and classroom management from the perspective of teachers, students and external observers: a multimethod approach. *Learning Environments Research*, *22*, 101–116. https://doi.org/10.1007/s10984-018-9269-x

Schweig, J., Hamilton, L. S., & Baker, G. (2019). *School and Classroom Climate Measures: Considerations for Use by State and Local Education Leaders* [Research Report N RR-4259-FCIM]. RAND Corporation. https://doi.org/10.7249/RR4259

Senden, B., Nilsen, T., & Teig, N. (2023). The validity of student ratings of teaching quality: Factorial structure, comparability, and the relation to achievement. *Studies in Educational Evaluation*, 78, 101274. https://doi.org/10.1016/j.stueduc.2023.101274

Smith-Crowe, K., Burke, M. J., Cohen, A., & Doveh, E. (2014). Statistical significance criteria for the rWG and average deviation interrater agreement indices. *Journal of Applied Psychology*, *99*(2), 239–261. https://psycnet.apa.org/doi/10.1037/a0034556

Stovner, R. B., Klette, K., & Nortvedt, G. A. (2021). The instructional situations in which mathematics teachers provide substantive feedback. *Educational Studies in Mathematics*, *108*(3), 533–551. https://doi.org/10.1007/s10649-021-10065-w

Tarkian, J., Lankes, E. M., & Thiel, F. (2019). Externe Evaluation - Konzeption und Implementation in den 16 Ländern [External evaluation – Conceptualizaion and implementation in the 16 federal states]. In F. Thiel, J. Tarkian, E.-M. Lankes, N. Maritzen, T. Rieke-Baulecke, & A. Kroupa (Eds.), *Datenbasierte Qualitätssicherung und -entwicklung in Schulen* (pp. 105–183). Springer. https://doi.org/10.1007/978-3-658-23240-5_5

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, *46*, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research*, *49*(2), 127–152. https://doi.org/10.1080/00131880701369651

Wemmer-Rogh, W., Grob, U., Charalambous, C. Y., & Praetorius, A. K. (2024). Measurement invariance between subjects: what can we learn about subject-related differences in teaching quality?. *ZDM*, 1–14. https://doi.org/10.1007/s11858-024-01622-7

White, M. C. (2022). A Validity Framework for the Design and Analysis of Studies Using Standardized Observation Systems. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of Analyzing Teaching Quality. Potentials and Pitfalls*. (pp. 89–120). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021-03

White, M., & Klette, K. (2023). What's in a score? Problematizing interpretations of observation scores. *Studies in Educational Evaluation*, *77*, 101238. https://doi.org/10.1016/j.stueduc.2023.101238

White, M., & Klette, K. (2024). Signal, error, or bias? exploring the uses of scores from observation systems. *Educational Assessment, Evaluation and Accountability*, 1–24. https://doi.org/10.1007/s11092-024-09427-8

# 4

# STUDY 2

Daltoè, T., Ruth-Herbein, E., Brucker, B., Jaekel, A.-K., Trautwein, U., Fauth, B., Gerjets, P., & Göllner, R. (2024). Immersive insights: Unveiling the impact of 360-degree videos on pre-service teachers' classroom observation experiences and teaching-quality ratings. *Computers & Education*, *213*, 104976. https://doi.org/10.1016/j.compedu.2023.104976

**Abstract**

Classroom videos are commonly used to observe and assess teaching quality in both teacher education and research on teaching and learning. In recent years, traditional video environments have increasingly been complemented by 360-degree videos, which promise a more immersive and realistic classroom experience and may affect the teaching-quality ratings that result. The aim of the present study was to explore differences between immersive 360-degree and traditional video environments in pre-service teachers' (PSTs') classroom observation experiences and teaching-quality ratings. Seventy-five PSTs observed two classroom videos: one using an immersive 360-degree video environment and one using a traditional video environment. For 360-degree videos, observers reported higher cognitive, affective, and physiological involvement in the classroom situation, higher motivation, and higher mental effort when making classroom observations. For one aspect of teaching quality (*focus on key concepts*), the observation-based ratings resulting from 360-degree videos were better aligned with experts' assessments of the videos. Furthermore, the results showed that the differences between video environments remained significant after the novelty of VR technology was controlled for.

## Introduction

Classroom videos are widely used as representations of teaching practice in teacher education and research. In teacher education, classroom videos are used, for example, to train teachers' professional competencies (Gaudin & Chaliès, 2015). In research, teaching-quality ratings in video studies have helped develop a deeper understanding of teaching and establish theories about teaching quality (e.g., Klieme et al., 2009). Classroom videos are typically presented on a standard personal computer (PC). In recent years, however, traditional classroom videos have increasingly been complemented by new technologies, such as 360-degree videos (e.g., Gold & Windscheid, 2020). In 360-degree videos, observers can freely choose their field of view of the classroom (Snelson & Hsu, 2020). Especially when presented as virtual reality (VR) videos, 360-degree videos promise a more immersive and authentic video and learning environment compared with traditional videos (Snelson & Hsu, 2020). 360-degree videos might have the potential to substantially move the current practice of video observation toward a more realistic classroom experience.

Observing 360-degree classroom videos in immersive environments is believed to offer clear advantages over traditional video environments on a standard PC, such as greater involvement in the classroom situation (e.g., Rupp et al., 2019) and increased motivation to observe the classroom and learn from classroom videos (e.g., Huang et al., 2021). It is reasonable to assume that a more realistic classroom observation in immersive video environments could also pose challenges for such observations. The increased complexity of the classroom observation with visual information surrounding the observer might lead to cognitive overload (e.g., Peterson et al., 2018), and observers might overlook relevant events in the video (Ardisara & Fung, 2018). Missing relevant classroom events might threaten the quality of teaching-quality ratings. Hence, before immersive classroom video environments are used in teacher training or research, it is crucial to understand how they affect classroom observations and teaching-quality ratings.

The first studies on 360-degree videos in teacher education have provided initial evidence that observers experience greater immersion and presence in the classroom situation than in traditional classroom videos (Ferdig & Kosko, 2020; Gold & Windscheid, 2020). However, little is known about how the use of immersive video environments affects teaching-quality ratings, especially for 360-degree classroom videos observed through VR glasses. Furthermore, the novelty effect that is evident for VR technology (Huang, 2020; Koch et al., 2018) might account for differences in classroom observations and teaching-quality ratings

observable for immersive 360-degree videos. For this reason, in the present study, we explored differences in classroom observation experiences and quantitative teaching-quality ratings between immersive and traditional video environments, while considering the novelty of wearing VR glasses for some participants as a moderator of differences between video types.

**Classroom Videos in Teacher Education and Research**

Video recordings of teaching capture complex classroom interactions and make them analyzable. Therefore, classroom videos have been established as a method for representing teaching in teacher education and research (e.g., Gaudin & Chaliès, 2015). In teacher education, videos are often used in university courses for training purposes (e.g., Santagata & Guarino, 2011). For instance, classroom videos were shown to be an appropriate method for training core practices of teaching (Grossman, 2021) or fostering professional competencies, such as professional vision of classroom management (Weber et al., 2018), feedback competence (Prilop et al., 2020), or reflective skills (Hamel & Viau-Guay, 2019). In research, classroom videos have usually been recorded and analyzed in large-scale video studies. Prominent examples of such video studies include the TIMSS video study (Hiebert & Stigler, 2000), the German-Swiss Pythagoras study (Klieme et al., 2009), and the TALIS video study (Ainley & Carstens, 2018). Research using video observations to assess teaching has made significant contributions to key findings about teaching and learning in recent decades (Klette, 2023). For example, systematic analyses of classroom videos have helped develop and test theories about *teaching quality*, which refers to the quality of certain teaching practices that are significant for student learning. These practices involve teachers' behavior and interactions between teachers and students in the classroom (Zee & Koomen, 2016). In video studies, researchers have identified three overarching dimensions of teaching quality—cognitive activation, student support, and classroom management—and established a framework for these dimensions (Klieme et al., 2009; Praetorius et al., 2018).

Despite the beneficial use of classroom videos, traditional classroom videos have limitations when used for teaching-quality assessments. Even though videos are used as a representation of practice, observing teaching in videos is far from observing teaching in an actual classroom. This difference might be particularly important for video-based trainings of teachers' competencies, as Klieme et al. (2007) argued that competencies are situation-specific. In order to acquire new competencies, the learning environment should therefore be as close as possible to the situation where the competency is needed. Another limitation concerns the fixed camera perspectives in traditional classroom videos. Unlike observations in real-life

classrooms, the perspective of traditional classroom videos on what is happening in the classroom is limited to the video camera's field of view, which is problematic for classroom observations, as prior research has shown significant effects of camera perspective on perceptions of teaching and resulting inferences about teaching quality (Cortina et al., 2018; Paulicke et al., 2019). Cortina et al. (2018) compared video recordings from the teacher's field of view with video recordings from the students' field of view and found more student-focused comments about the teaching for the video recordings from the teacher's perspective. Moreover, Paulicke et al. (2019) investigated differences in teaching-quality ratings for the same lessons recorded from different camera angles. Their results showed significantly different teaching-quality ratings for the different camera angles for all dimensions of teaching quality. Furthermore, individual students' behaviors were perceived primarily from the camera using the students' perspective (Paulicke et al., 2019). As teaching quality is understood as a co-construction between students and teachers (e.g., Fauth et al., 2020), ratings should be based on observations of activities across the entire classroom.

Taken together, the limitations of traditional classroom videos make it clear that a more realistic form of video observation could (a) aid teacher competency training by better resembling an actual teaching situation and (b) improve the validity of teaching-quality assessments. One promising way to improve the realism of classroom videos is to use 360-degree videos.

**Three-Hundred-Sixty-Degree Classroom Videos**

Three-hundred-sixty-degree videos, also known as immersive videos, are an emerging technology in education (Ranieri et al., 2022; Snelson & Hsu, 2020). Immersion in the context of 360-degree videos refers to the extent to which the technology creates a comprehensive, intense, surrounding, and vivid illusion of a virtual environment for the observer (Slater & Wilbur, 1997). To provide this immersive environment, 360-degree cameras capture the surroundings in all directions, thus allowing the observer to freely choose their field of view on the video content from various positions of the 360-degree camera in the classroom. Choosing the field of view works, for example, by scrolling with a mouse (in 360-degree PC videos) or turning one's head (in 360-degree VR videos).

Three-hundred-sixty-degree classroom videos seem to offer a more realistic classroom experience, which might come with benefits but also increased complexity and challenges for observers. One central benefit of 360-degree videos is the flexible field of view, which allows the observer to choose their perspective on the classroom situation (Balzaretti et al., 2019). With

regard to the observation of classroom videos, Gold and Windscheid (2020) pointed out that, via free choice of perspective on the classroom situation, observers are able to obtain more information about the classroom situation, potentially increasing the validity of teaching-quality ratings. However, the flexible viewing perspective in 360-degree videos might also lead observers to overlook relevant events (Ardisara & Fung, 2018), thereby potentially threatening the validity of teaching-quality ratings. Another advantage of immersive video environments could be an increase in motivation from the video content (e.g., Makransky & Peterson, 2021). For example, situational interest promotes attention and engagement in a learning task (Harackiewicz et al., 2016). This effect might also be evident for the task of classroom observation. Increased attention to the classroom video might positively affect teaching-quality ratings. Another benefit of using immersive 360-degree videos is that they evoke stronger feelings of involvement in the classroom situation (Huang et al., 2021; Rupp et al., 2019). Previous research has assessed immersive experiences along multiple dimensions, including cognitive, affective, and physical involvement (e.g., Zachrich et al., 2020), all of which can be assumed to become relevant when observing a classroom situation in traditional or immersive video environments. Strong feelings of involvement in the immersive video environment could impact observers' attentional capacities (Makransky & Peterson, 2021), thus potentially leading to a different assessment of teaching quality in classroom videos. However, if the required attentional capacity is too high, observers might experience cognitive overload and might not be able to focus on the relevant classroom events, potentially threatening the validity of the resulting teaching-quality ratings. For example, studies from medical contexts found that the use of immersive VR was associated with increased cognitive load and reduced task performance (Frederiksen et al., 2020; Peterson et al., 2018).

The use of immersive 360-degree videos in teacher education and research is an emerging field of research (Atal et al., 2023; Roche et al., 2023). Kunz and Zinn (2022) found that pre-service teachers (PSTs) experienced high technology acceptance and feelings of presence and immersion in 360-degree classroom videos (Kunz & Zinn, 2022). Other studies have provided evidence that feelings of presence and immersion are significantly higher than in traditional classroom videos (Ferdig & Kosko, 2020; Gold & Windscheid, 2020). Regarding emotions and workload experienced during classroom observation and resulting teaching-quality ratings, Gold and Windscheid (2020) found no differences between 360-degree classroom videos and traditional classroom videos. However, in their study, PSTs observed both types of videos on standard PCs, and the authors argued that differences might occur if more immersive devices (e.g., VR glasses) were used to watch the 360-degree videos (Gold &

Windscheid, 2020). Walshe and Driver (2019) used 360-degree classroom videos to teach PSTs to reflect on teaching situations. They found that reflecting on 360-degree classroom videos promoted a deeper and more sophisticated understanding of microteaching practices. Furthermore, they found increases in PSTs' teaching-related self-efficacy (Walshe & Driver, 2019). Kosko et al. (2021) compared three different types of classroom videos: traditional videos observed on a PC, 360-degree videos observed on a PC, and 360-degree videos observed through VR glasses. They assessed written responses of what PSTs noticed when viewing the classroom videos as well as screen recordings and found that the PSTs in both 360-degree video conditions focused more on students' actions in their written comments than the traditional classroom video group did (Kosko et al., 2021). Moreover, PSTs observing the 360-degree videos through VR glasses looked at different classroom areas and described the classroom events with more specificity than PSTs observing the 360-degree videos on a PC (Kosko et al., 2021). Another study by Kosko et al. (2022) focused on PSTs' noticing in 360-degree videos in VR and found that successful and detailed noticing of relevant classroom events depended on how observers positioned the students and the teacher in their field of view. All in all, prior studies suggest positive effects of 360-degree classroom videos, especially on classroom observation experiences (Ferdig & Kosko, 2020; Gold & Windscheid, 2020; Kunz & Zinn, 2022; Walshe & Driver, 2019). However, it is possible that these positive experiences are associated with the novelty of using a new technology. The *novelty effect* occurs when users are more motivated or perceive higher usability for a technology that is new to them (Huang, 2020; Koch et al., 2018). When observers experience immersive video technology for the first time, the novelty effect might account for any differences that occur between immersive and traditional classroom videos. So far, the novelty effect has not been considered in research on 360-degree classroom videos.

**Present Study**

The use of 360-degree videos and VR in teacher education has a great deal of potential. Most importantly, previous research has demonstrated that observers feel more immersed in 360-degree classroom videos, indicating that 360-degree videos provide a highly realistic classroom scenario. On the other hand, until now, less attention has been paid to determining how immersive video environments in VR affect raters psychologically, regarding their classroom observation experiences and teaching-quality ratings that result from their classroom observations. Furthermore, more research is needed on the novelty effect in classroom video research. To fill this research gap, we conducted a controlled study to compare PSTs' classroom

observation experiences and the resulting teaching-quality ratings between immersive and traditional video environments. With Research Questions 1-3 (RQ1-RQ3), we investigated different experiences related to classroom observations.

> **RQ1:** How do PSTs' experiences of cognitive, affective, and physiological involvement in the classroom differ between classroom observations from immersive 360-degree classroom videos and traditional classroom videos?

On the basis of previous studies on 360-degree videos (Breves & Schramm, 2021; Ferdig & Kosko, 2020; Gold & Windscheid, 2020; Huang et al., 2021; Rupp et al., 2019), we expected higher cognitive, affective, and physiological involvement in the classroom situation in the immersive video environment.

In addition, we investigated how motivating the classroom observations were for observers in the immersive and traditional video environments.

> **RQ2:** How does the extent to which PSTs feel motivated from watching the classroom videos differ between classroom observations from immersive 360-degree classroom videos and traditional classroom videos?

Studies on learning with immersive VR have consistently found increases in enjoyment and intrinsic motivation in VR conditions compared with learning with less immersive media (Huang et al., 2021; Liu et al., 2022; Makransky & Lilleholt, 2018; Meyer et al., 2019). For 360-degree videos, Rupp et al. (2019) found a positive relationship between the feeling of presence and the desire to learn about the video's subject matter. On the basis of these findings, we expected that PSTs would be more motivated by the classroom videos presented in the immersive video environment.

Gold and Windscheid's (2020) study found no significant differences in perceived workload from classroom observations of 360-degree and traditional classroom videos. However, in their study, participants observed 360-degree videos on PCs. In other contexts, the use of VR was associated with increased cognitive load (Frederiksen et al., 2020; Peterson et al., 2018). This research question needs to be investigated further in the context of classroom observation with more immersive devices (e.g., VR glasses; Kosko et al., 2021). For this reason, we compared the cognitive capacity required to make classroom observations (mental load and mental effort; Krell, 2017) between the immersive and traditional video environments.

**RQ3:** How does the cognitive capacity PSTs have available to observe the classroom differ between observations of immersive 360-degree classroom videos and traditional classroom videos?

In teacher education and research, it is common for (pre-service) teachers to assess aspects of teaching quality from classroom videos (e.g., Gaudin & Chaliès, 2015; Klieme et al., 2009). For this reason, with RQ4, we investigated the teaching-quality ratings resulting from the classroom observations made in the different video environments.

**RQ4:** How do PSTs' teaching-quality ratings (absolute ratings, accuracy of ratings) differ between classroom observations from immersive 360-degree classroom videos and traditional classroom videos?

When researching the effects of immersive video environments on classroom observations, the novelty of this technology might be a confounding factor (Huang, 2020; Koch et al., 2018). For this reason, we investigated whether potential differences found from addressing RQ1-RQ4 differed between observers with and without prior VR experience.

**RQ5:** Is the novelty of wearing VR glasses for specific participants associated with potential differences in their subjective classroom observation experiences (involvement, motivation, cognitive capacity) and teaching-quality ratings between immersive 360-degree classroom videos and traditional classroom videos?

## Method

The present study is part of an interdisciplinary project on the potential of VR in teacher education and teaching-quality development by the University of Tübingen, the Universities of Education Freiburg and Heidelberg, and the Institute for Educational Analysis Baden-Württemberg. The data collection was preregistered on aspredicted.org (https://aspredicted.org/67J_93Q). This study partially addresses the preregistered research questions; however, we added an examination of the novelty effect as an exploratory analysis.

### Design and Participants

We used a within-person research design to compare PSTs' observation experiences and the resulting teaching-quality ratings between immersive and traditional video environments. Each participant observed two classroom videos: one video in an immersive 360-degree video environment using VR glasses and one video in a traditional video environment on a standard PC. The video content and order of the video conditions was randomly assigned to the participants.

The study was conducted in a lab at the Leibniz-Institut für Wissensmedien in Tübingen in the winter term between December 2022 and February 2023. A total of $N = 75$ PSTs (72% female, 28% male, $M_{age} = 24.72$, $SD_{age} = 2.02$) from different subjects from two universities participated in the study. The majority of the participants ($N = 71$; 95%) were master's students, and some ($N = 4$; 5%) were bachelor's students. Half of the participants ($N = 37$; 49.2%) had prior experience wearing VR glasses, whereas VR glasses were completely new for the other half ($N = 38$; 51.8%).

We recruited participants from the university mailing list, different subject-specific mailing lists for PSTs, and advertisements in teaching courses for PSTs in the master's program. Participation in the study was aimed at PSTs in the master's program, as participants should already have prior teaching and classroom observation experience. The participants received an incentive of 20€.

**Video Material**

We used five mathematics classroom videos from two different classrooms from schools in the state of Baden-Württemberg, Germany. The classroom videos show staged teaching situations. Staged videos allow the direct illustration of crucial teaching events and critical situations (Codreanu et al., 2020; Piwowar et al., 2018; Seidel et al., 2022). The videos used in this study depict different quality gradations of three specific teaching-quality aspects from a classroom observation system designed for teaching-quality development in Baden-Württemberg, Germany (Ruth-Herbein et al., 2022). This classroom observation system contains 11 items assessing aspects of the three basic dimensions of teaching quality: cognitive activation, student support, and classroom management (Klieme et al., 2009; Praetorius et al., 2018). The classroom videos showed positive and negative examples of the following three teaching-quality aspects: *focus on key concepts* (cognitive activation), *challenging tasks* (cognitive activation), and *classroom disruptions* (classroom management). The scripts for the staged classroom videos were developed by a team of researchers and practitioners consisting of mathematics didactics experts, teaching-quality experts, and teachers. Therefore, expert assessments of the teaching quality in the staged classroom videos were available. Table 1 provides an overview of the five classroom videos, the video content, and the quality gradation for the focused teaching-quality aspects (*focus on key concepts*, *challenging tasks*, *classroom disruptions*) as intended by the experts who developed the staged classroom videos.

The classroom situations were simultaneously recorded as traditional classroom videos and 360-degree classroom videos. The traditional classroom videos were videotaped with two

*Blackmagic Pocket Cinema* cameras from two camera perspectives that are commonly used for classroom videos: one teacher-focused perspective recorded from the back of the classroom and one student-focused perspective recorded from the front corner (e.g., Jacobs et al., 2003). We edited the final classroom videos as split-screen videos showing the two camera perspectives simultaneously to make sure that all important classroom events were visible (Kilburn, 2014; Windscheid & Will, 2018). The 360-degree classroom videos were videotaped with one *Insta360 ONE X* camera. The 360-degree camera was placed in the middle of the classroom. From this point of view, all important classroom events were visible. With the chosen camera perspectives for the two types of videos, we aimed to best approximate common practices for classroom videos (e.g., Hofman, 2022; Kilburn, 2014; Kosko et al., 2021). Figure 1 shows a screenshot from each of the two video conditions.

**Figure 1**

*Screenshots From the Traditional Video and Three Different Angles of View in the 360-Degree Video*

**Table 1**

*Overview of the Classroom Videos Used in the Study*

| | Duration | Grade | Teaching content | Scripted teaching quality |
|---|---|---|---|---|
| Video 1 | 6:54 min. | Eighth grade | Distributive law (algebra) | · Focus on key concepts: High quality<br>· Classroom disruptions: Low quality |
| Video 2 | 7:22 min. | | | · Focus on key concepts: High quality<br>· Classroom disruptions: High quality |
| Video 3 | 6:33 min. | | | · Focus on key concepts: Low quality<br>· Classroom disruptions: High quality |
| Video 4 | 9:00 min. | 10th grade | Calculation of the area of a circle sector (geometry) | · Challenging tasks: High quality |
| Video 5 | 9:06 min. | | | · Challenging tasks: Low quality |

*Note.* Videos 1, 2, and 3 and Videos 4 and 5 were each recorded in the same class with the same teacher. The teaching content and grade are the same for Videos 1, 2, and 3 and Videos 4 and 5.

**Instruments**

We used standardized questionnaires to investigate the constructs we identified as relevant for the classroom observation experience. Here, we used previously created and validated scales (e.g., Gold & Windscheid, 2020; Hajahmadi & Marfia, 2023). To examine the different facets of involvement and ensure a detailed assessment, we assigned two scales each to the facets of cognitive, affective, and physiological involvement. We assessed cognitive involvement with the scales *immersion* (Immersive Experience Questionnaire for Film and TV [Film IEQ]; Rigby et al., 2019; 1 [*disagree*] to 7 [*agree*]) and *focus of attention* (adapted from Zachrich et al., 2020; 1 [*disagree*] to 5 [*agree*]). Affective involvement was assessed with the scales *compassion for the teacher* and *compassion for the students* (both adapted from Zachrich et al., 2020; 1 [*disagree*] to 5 [*agree*]). To investigate physiological presence, we used two scales from Wirth et al.'s (2008) Measurement, Effects, Conditions of Spatial Presence Questionnaire: *self-location* and *possible actions* (1 [*disagree*] to 5 [*agree*]). To investigate the extent to which participants felt motivated to work with classroom videos, we developed our own scale (1 [*disagree*] to 5 [*agree*]), as no scales have been established for this specific context, although Seidel et al. (2011) used a similar scale. To investigate cognitive capacity to observe the classroom, we used two task-related scales adapted from Krell (2017): *mental load* and *mental effort* (1 [*disagree*] to 7 [*agree*]). Mental load describes the cognitive capacity required to process the complexity of a task (e.g., the classroom observation). Mental load is connected to mental effort, the cognitive capacity an individual invests in the task (Krell, 2017). Table 2 presents detailed information about the questionnaires, including the scales and sample items. Table 3 provides additional information about internal consistencies as a measure of a scale's reliability. Reliability was satisfactory for all scales, with the exception of the *immersion* scale in the VR condition (see Table 3). The novelty of wearing VR glasses investigated in RQ5 was assessed with a single item ("Have you used VR glasses in other contexts before?") with a dichotomous response format (yes/no). For the teaching-quality ratings, we used the standardized classroom observation system by Fauth et al. (2022) comprising 11 individual teaching-quality items (1 [*disagree*] to 4 [*agree*]). Three of the teaching-quality aspects included in the observation system by Fauth et al. (2022) provided the theoretical foundation for the staged video production: *focus on key concepts*, *challenging tasks*, and *classroom disruptions*. As these three teaching-quality items are of special interest for the teaching-quality assessment of the staged classroom videos, we focus on these items in the following. In addition to the quantitative teaching-quality items, participants were given the chance to explain their teaching-quality ratings in open responses.

**Table 2**

*Overview of the Scales Assessed With Sample Items and the Three Teaching-Quality Items*

| Scale | | Number of items | Sample item |
|---|---|---|---|
| Cognitive involvement | Immersion | 4 | I was focused on the VR experience/ the classroom video on the PC. |
| | Focus of attention | 4 | I was fully focused on the lesson. |
| Affective involvement | Compassion for teacher | 3 | I empathized with the teacher during the lesson. |
| | Compassion for students | 3 | I empathized with the students during the lesson. |
| Physiological involvement | Self-location | 3 | I felt as though I was present in the classroom myself. |
| | Possible actions | 3 | I had the impression that I could take action myself in the classroom. |
| Motivated by the classroom video | | 3 | Observing the classroom video was interesting. |
| Cognitive capacity to observe the classroom | Mental load | 6 | It was challenging to follow what was going on in class. |
| | Mental effort | 6 | I put effort into observing the classroom. |
| Teaching quality | Focus on key concepts | 1 | The lesson focused on the key concepts that students are expected to understand. |
| | Challenging tasks | 1 | The teacher used tasks and questions that challenged students' higher order thinking. |
| | Classroom disruptions | 1 | The lesson ran for the most part without disruption. |

**Procedure**

After arriving at the lab, participants gave their informed consent. They were then fit with an Empatica E4 wristband for collecting physiological data. To familiarize participants with the observation system used for the teaching-quality ratings in this study, they watched a brief introductory video to the classroom observation system (duration: 08:46 min) on a standard PC screen. After the introductory video, we instructed them to observe the first classroom video carefully and to assess the quality of the teaching afterwards. Then, participants observed the first video either as a traditional video on a standard PC (SMI laptop) with headphones or as a 360-degree video through VR glasses (HP Reverb). After they observed the first video, participants answered a questionnaire on their classroom observation experience and gave their teaching-quality ratings (see 2.3). In the immersive 360-degree video condition, the first part of the questionnaire (cognitive, affective, and physiological involvement) appeared in the immersive environment directly and was answered by using a controller. The rest of the questionnaire (including the teaching-quality ratings) was answered in the online-survey platform *Unipark*. In the traditional video condition, all questions were answered in the online questionnaire. However, for the first part of the questions, a screenshot of the classroom was presented so that the questionnaire conditions would be as comparable as possible between the two video conditions. After completing the questionnaire for the first classroom video, participants switched seats and observed the second classroom video in the other condition they had not previously experienced (immersive or traditional video, respectively). After observing the second video, participants completed the same online questionnaire for the second classroom video. Figure 2 illustrates the two video conditions each participant experienced in the study. The total duration of the experiment was approximately 90 min.

**Figure 2**

*Two Conditions: Traditional Video Environment (Left) and Immersive 360-Degree Video Environment (Right)*

## Statistical Analysis

We tested for differences between immersive and traditional video environments regarding different outcome domains in RQ1-RQ3 (e.g., involvement, motivation, cognitive capacity). Each domain comprised a set of different measures. To analyze these study outcomes, we conducted two separate analyses. First, we employed multivariate tests to examine differences between immersive and traditional video environments for each of the outcome domains. Specifically, due to the within-subject design of our study, we computed a repeated-measures MANOVA. MANOVA enabled us to simultaneously assess the variables within each outcome domain and test for multivariate group differences. MANOVA yields a more reliable level of Type I errors on multiple response measurements than a set of separate analyses, and it offers greater efficiency to reliably detect group differences (Stevens, 2012). Second, we performed univariate tests for each individual variable to identify the specific variables on which the groups potentially differed. Once again, we accounted for the within-subject design by using repeated-measures analyses for all univariate tests. To facilitate result interpretation, we calculated $\eta^2$, where values of .01-.06 indicate small effects, values of .06-.14 indicate moderate effects, and values of .14 and above indicate large effects (Cohen, 1988).

Regarding teaching-quality ratings (RQ4), we were interested in whether participants systematically gave more positive or more negative ratings of teaching quality in one of the video environments. Therefore, we calculated a repeated-measures MANOVA to see if there were within-person differences between the video environments in the absolute ratings. In a second step, we were interested in whether one of the video environments resulted in more accurate teaching-quality ratings in terms of the manipulation of teaching quality in the staged classroom videos. Therefore, we conduced univariate tests for the three teaching-quality items. In this context, we controlled for the focused quality aspect and its valence (e.g., positive or negative manipulation) and modeled additional interactions between the video environment and the teaching-quality manipulation as well as the valence of the manipulations. The resulting interactions provided information about how the rating was affected by the video environment for the teaching-quality manipulations and the valence of the manipulation (two-way interactions). Furthermore, we modeled one three-way interaction between video environment, teaching-quality manipulation, and valence of the manipulation in order to test whether the video environments resulted in more accurate teaching-quality ratings, meaning that the ratings were more in the direction of manipulation for a specific aspect of quality. As participants rated teaching in the content area of mathematics, we included an additional variable to control for participants' specialization (mathematics vs. not mathematics).

With RQ5, we asked whether the novelty of wearing VR glasses was associated with potential differences between the video environments in RQ1-RQ4. Therefore, we recalculated the analyses we conducted to address the previous research questions but included the novelty variable.

We analyzed the data with the software IBM SPSS Statistics (Version 27; IBM Corporation, 2020). The alpha level was set at .05 in accordance with common practice.

## Results

### Descriptive Results

Descriptive statistics for the scales used to measure classroom observation experiences and the three teaching-quality items are presented in Table 3. Whereas Cronbach's α is reported as an indicator of the internal consistency of the scales, we report interrater agreement as a measure of consistency in the teaching-quality ratings in the different video environments. We analyzed interrater agreement between raters with the average absolute deviation index ($AD_M$; Burke et al., 1999). The $AD_M$ describes the mean absolute deviation of a group of raters from the group mean of all raters. Consequently, lower values indicate higher interrater agreement. The agreement was higher in the immersive video environment for the teaching-quality items *focus on key concepts* and *classroom disruptions* and higher in the traditional video environment for the teaching-quality item *challenging tasks* (see Table 3).

### Differences in the Classroom Observation Experience

With RQ1, we asked whether PSTs reported different cognitive, affective, and physiological involvement in the classroom situation in immersive 360-degree versus traditional video environments. With the multivariate test, we found that, overall, participants reported significantly higher involvement in the classroom situation in the immersive video environment than in the traditional video environment with a large effect size ($p < .001$, $\eta^2 = .87$, see Table 4). When taking a look at the individual variables indicating cognitive, affective, and physiological involvement in the univariate tests, we found that the difference was significant for all the variables with moderate to large effect sizes, with the exception of *compassion for the teacher* (see Table A1 in Appendix).

To address RQ2, we investigated differences in the extent to which the classroom video was motivating between the immersive and traditional video environments. We found that the motivation reported from the classroom video was significantly higher in the immersive video environment with a large effect size ($p < .001$, $\eta^2 = .28$, see Table 4).

With RQ3, we investigated whether observers reported different levels of cognitive capacity to observe the classroom in immersive and traditional video environments, measured via mental load and mental effort. For this research question, the multivariate test showed significantly higher values for the classroom observation in the immersive video environment with a moderate effect size ($p = .01$, $\eta^2 = .13$, see Table 4). However, for the single variables in the univariate tests, we found that only mental effort was significantly higher for the classroom

observation in the immersive video environment with a moderate effect size ($p = .001$, $\eta^2 = .13$), but no differences in mental load occurred between the two types of videos ($p = .40$, see Table A1 in Appendix).

**Differences in Teaching-Quality Ratings**

With RQ4, we investigated whether the absolute teaching-quality ratings resulting from the classroom observation differed between the immersive and traditional video environments. We found no significant differences in the observation ratings for the multivariate test in which all ratings were combined ($p = .725$, see Table 4). The univariate tests revealed that for the aspects *focus on key concepts* and *challenging tasks*, the ratings partially differed between video environments (e.g., when scripted videos were positively or negatively manipulated but irrespective of the specific teaching-quality aspect the videos focused on). Most importantly, however, the ratings regarding *focus on key concepts* resulting from the immersive video environment were more positive for the classroom videos that were scripted as positive regarding the aspect *focus on key concepts* (three-way interaction) compared with all other classroom videos and compared with the ratings resulting from the traditional video environment ($b = 0.27$, $SE = .13$, $p = .040$, $\eta^2 = .01$, see Table 5). This value indicates that, for this specific teaching-quality aspect, the ratings showed better agreement with the experts who developed the scripted classroom videos. We did not find differences between participants who specialized in the content area of mathematics and participants with other specializations. According to Cohen's effect-size classifications, the differences between video environments in teaching-quality ratings can be described as small (Cohen, 1988).

**Table 3**

*Overview of the Descriptive Statistics for the Scales, Presented Separately for the Standard Computer (PC) and the 360-Degree VR Condition (VR)*

| Scale | | *M* | | *SD* | | Cronbach's $\alpha$ | |
|---|---|---|---|---|---|---|---|
| | | PC | VR | PC | VR | PC | VR |
| Cognitive involvement | Immersion | 5.27 | 6.28 | 1.10 | 0.73 | .80 | .64 |
| | Focus of attention | 4.04 | 4.26 | 0.79 | .63 | .80 | .76 |
| Affective involvement | Compassion for teacher | 2.95 | 2.84 | 1.01 | 1.08 | .89 | .92 |
| | Compassion for students | 2.86 | 3.45 | 0.91 | 1.00 | .80 | .85 |
| Physiological involvement | Self-location | 1.74 | 4.20 | 0.79 | 0.86 | .84 | .82 |
| | Possible actions | 1.28 | 2.87 | 0.60 | 1.25 | .91 | .90 |
| Motivated by the classroom video | | 3.53 | 4.26 | 0.92 | 0.90 | .84 | .84 |
| Cognitive capacity to observe the classroom | Mental load | 2.72 | 2.53 | 1.21 | 1.36 | .90 | .94 |
| | Mental effort | 5.75 | 6.06 | 0.99 | 0.86 | .85 | .81 |
| | | *M* | | *SD* | | $AD_M$ | |
| | | PC | VR | PC | VR | PC | VR |
| Teaching quality | Focus on key concepts | 3.52 | 3.61 | 0.70 | 0.54 | 0.49 | 0.46 |
| | Challenging tasks | 2.89 | 2.87 | 0.89 | 0.95 | 0.68 | 0.69 |
| | Classroom disruptions | 3.49 | 3.59 | 1.00 | 0.93 | 0.23 | 0.17 |

*Note.* $AD_M$ = Average absolute deviation index.

**Table 4**

*Multivariate Main Effects of Video Environment for Aggregated Outcome Variables*

| Within-subjects effect | Outcome | | Value | *F* | Hypothesis *df* | Error *df* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|---|---|
| Video environment | Cognitive, affective, and physiological involvement | Wilk's lambda | .13 | 79.29 | 6 | 69 | < .001 | .87 |
| | Motivated by the classroom video | | .73 | 28.06 | 1 | 74 | < .001 | .27 |
| | Cognitive capacity to observe the classroom | | .87 | 5.58 | 2 | 73 | .006 | .13 |
| | Teaching quality | | .98 | 0.44 | 3 | 72 | .725 | .02 |

**Table 5**

*Univariate Tests of Differences in Teaching-Quality Ratings*

| | Focus on key concepts | | | | Challenging tasks | | | | Classroom disruptions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimate | SE | *p* | $\eta^2$ | Estimate | SE | *p* | $\eta^2$ | Estimate | SE | *p* | $\eta^2$ |
| Video Environment | 0.15 | .05 | .001 | .01 | 0.16 | .11 | .141 | .01 | 0.01 | .05 | .874 | <.01 |
| Video Environment×Valence | -0.17 | .08 | .031 | .01 | -0.38 | .18 | .034 | .03 | -0.01 | .07 | .918 | <.01 |
| Video Environment× Manipulation of Respective Teaching-Quality Aspect | -0.23 | .09 | .012 | .01 | -0.22 | .17 | .196 | .01 | 0.05 | .08 | .563 | <.01 |
| Video Environment× Manipulation of Respective Teaching-Quality Aspect× Valence | 0.27 | .13 | .040 | .01 | 0.42 | .25 | .088 | .02 | -0.03 | .12 | .819 | <.01 |
| Video Environment×Subject | -0.12 | .08 | .157 | <.01 | -0.19 | .17 | .266 | <.01 | 0.05 | .09 | .537 | <.01 |

**Effects of the Novelty of Wearing VR Glasses**

With RQ5, we investigated whether the novelty of wearing VR glasses for some participants was associated with the differences in classroom observation experiences found in RQ1-RQ4. When we added the novelty variable, we found no main effect of novelty and no significant interactions between novelty and the video environment for either classroom observation experiences or teaching-quality ratings (see Table 6).

When we tested the associations between novelty and the individual variables in the univariate tests, we found two moderately sized significant main effects of novelty for the variables *compassion for students* ($p = .017$, $\eta^2 = .08$) and *self-location* ($p = .022$, $\eta^2 = .07$). We also found one moderately sized significant interaction between novelty and the video environment for immersion in the classroom situation ($p = .03$, $\eta^2 = .06$, see Table A2 in Appendix). The significant main effects indicate that observers without prior VR experience reported greater compassion for the students in the videos and greater physical self-location in the classroom in both video environments. More interestingly, the significant interaction effect indicates that, for observers without prior VR experience, the increase in immersion in the immersive video environment was significantly higher than for observers with prior VR experience (see Figure 3). However, all the differences between video environments remained significant after novelty was controlled for.

**Figure 3**

*Significant Video Environment × Novelty Interaction*

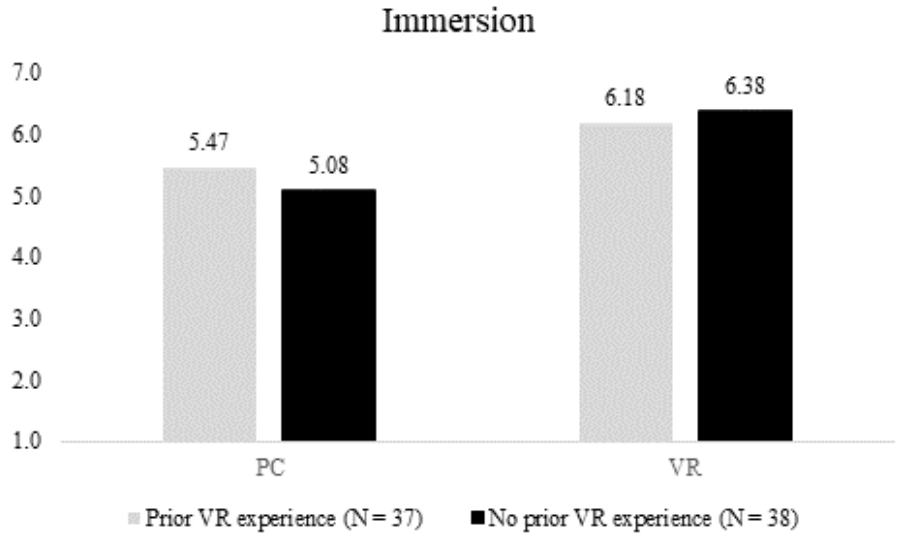


**Table 6**

*Multivariate Main Effects of Novelty and Novelty × Video Environment Interactions for Aggregated Outcome Variables*

| | | Outcome | | Value | *F* | Hypothesis *df* | Error *df* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Between-subjects effect | Novelty | Cognitive, affective, and physiological involvement | Wilk's lambda | 0.85 | 2.01 | 6 | 68 | .076 | .15 |
| | | Motivated by the classroom video | | 1.13 | 1.22 | 1 | 73 | .273 | .02 |
| | | Cognitive capacity to observe the classroom | | 1.00 | 0.10 | 2 | 72 | .905 | .00 |
| | | Teaching quality | | 0.99 | 0.20 | 3 | 71 | .899 | .01 |
| Within-subjects effect | Novelty×Video Environment | Cognitive, affective, and physiological involvement | Wilk's lambda | 0.84 | 2.24 | 6 | 68 | .050 | .17 |
| | | Motivated by the classroom video | | 0.96 | 2.87 | 1 | 73 | .094 | .04 |
| | | Cognitive capacity to observe the classroom | | 0.98 | 0.80 | 2 | 72 | .452 | .02 |
| | | Teaching quality | | 1.00 | 0.11 | 3 | 71 | .955 | .00 |

## Discussion

Deriving conclusions about teaching quality from observing classroom videos is a common method in both teacher education and research on teaching and learning (Gaudin & Chaliès, 2015; Praetorius et al., 2012). In this context, immersive 360-degree videos promise more realistic classroom observation experiences than traditional videos, but several opportunities and challenges come with this realism. The aim of the present study was to explore differences between immersive and traditional video environments in PSTs' classroom observation experiences and the resulting teaching-quality ratings and to additionally consider a potential novelty effect for participants wearing VR glasses for the first time.

We found that the observers felt more cognitively, affectively, and physically involved in the classroom situation in the immersive video environment (RQ1). These findings replicate previous research that found increased immersion and presence as indicators of involvement in 360-degree classroom videos compared with traditional classroom videos (Ferdig & Kosko, 2020; Gold & Windscheid, 2020). Additionally, our findings extend previous research by adding more facets of observation-based experiences. However, one of the scales (immersion in VR) showed a questionable internal consistency of $\alpha = .64$. This questionable internal consistency reflects the novelty of this research area. However, we believe that the questionable internal consistency of the scale had only a minor impact on our results, as a lower internal consistency will typically lead to a lower probability of finding any statistically significant effects. In the present study, all tests involving the immersion in VR scale showed practical differences, which were all statistically significant. In RQ2, we added a motivational variable to the research on 360-degree classroom videos and found that participants felt more motivated by the classroom video when they observed it in the immersive video environment. This finding is in line with previous research from other disciplines that found that using VR technology increased motivation compared with less immersive media (e.g., Makransky & Lilleholt, 2018; Meyer et al., 2019) or that the feeling of presence in 360-degree videos was associated with the desire to learn about the video content (Rupp et al., 2019). Furthermore, we found that PSTs reported greater cognitive capacity to observe the classroom in the immersive video environment (RQ3). Our findings showed that these differences were mainly related to experienced mental effort, thus indicating that observers invested more effort in observing the classroom in the immersive video environment. The greater effort put toward the observation task could be explained by the fact that observers had to look around to successfully observe the whole classroom. However, the participants did not perceive an increase in the complexity

of their task to observe the classroom situation (mental load; Krell, 2017). This result regarding mental load is consistent with the study by Gold and Windscheid (2020), who also found that participants did not experience an increased workload when watching 360-degree classroom videos compared with traditional videos. Although the 360-degree videos in Gold and Windscheid's (2020) study were observed on a standard PC, we were able to confirm the finding of no increase in mental load even for a more immersive environment in 360-degree videos when wearing VR glasses. Participants' cognitive capacity to observe a classroom in an immersive video environment might be comparable to their capacity to observe a real-life classroom.

Regarding teaching-quality ratings, immersive classroom observation did not result in a more positive or negative rating compared with the traditional observation setting (RQ4). In addition, the univariate tests revealed a difference between the two video environments with respect to the question of whether ratings of individual teaching-quality aspects were comparable to experts' ratings. We found that participants' ratings of the teacher's *focus on key concepts* were more in line with expert ratings in 360-degree classroom videos than in traditional videos. The results for the teaching-quality aspect *challenging tasks* were generally in the same direction, whereas the differences were not statistically significant for *classroom disruptions*. These findings might be explained by the extent to which the teaching referred to an interplay between the teacher and the students. Whereas the extent of classroom disruptions could be identified by focusing on the students only, the teacher's *focus on key concepts* and *challenging tasks* were more directly visible in the discourse in the classroom. Our results might suggest that the interplay between teacher and students is more salient in 360-degree classroom videos than in traditional classroom videos.

Furthermore, when innovative technologies have more positive effects than traditional approaches, the novelty of using the new technology may account for these differences (Huang, 2020; Koch et al., 2018). Investigating a possible novelty effect in RQ5, we found that wearing VR glasses for the first time was not associated with the differences between video environments investigated in RQ1-RQ4. However, for one aspect of cognitive involvement (immersion), we found that the difference was significantly higher for participants without prior VR experience. This effect suggests that the captivating effect of 360-degree videos in VR was particularly large for individuals with no prior VR experience. Nevertheless, the differences revealed between immersive and traditional video environments could not be explained by the technology's novelty for the user.

In summary, our results suggest that immersive 360-degree classroom videos can benefit classroom observation via increased engagement, increased motivation, and increased mental effort. In terms of the resulting teaching-quality ratings, we found less consistent differences between the two video environments. The extent to which the investigated teaching-quality aspects are embedded in teacher-student interactions might contribute to the differences we found between the video conditions but require further exploration in studies applying more systematic variation in teaching-quality aspects in videos. Overall, the findings suggest that 360-degree classroom videos provide a more realistic representation of classrooms, thus emphasizing the potential of technology in the context of teacher education and teachers' professional development.

**Limitations**

The present study extends the current state of research on the use of 360-degree classroom videos in teacher education and research. However, our study has several limitations. A key limitation is that the two video conditions differed in several ways. The traditional and 360-degree videos were recorded from different camera positions in the classroom. Furthermore, the potential observation angles differed: Whereas the traditional video was presented as a split-screen video with two fixed viewing axes from two perspectives, the 360-degree video was presented from one perspective without a fixed observation angle. In addition, observing the videos on different devices (PC vs. VR glasses) limited the comparability of the two conditions. The device and the degree of immersion it induces may affect the use of 360-degree videos (Rupp et al., 2019). To clearly distinguish the effects of video type and device, it would be necessary to integrate multiple conditions into a study: traditional video on PC, 360-degree video on PC, traditional video observed through VR glasses, 360-degree video observed through VR glasses (Rupp et al., 2019; Snelson & Hsu, 2020). These conditions would in turn require larger samples. Consequently, the small sample size, which limited the possible analyses, was another limitation of this study, although the sample size of $N = 75$ PSTs was larger than previous studies in the research field on 360-degree classroom videos. Another limitation of the present study is that the analyses of teaching-quality ratings were restricted to the teaching-quality aspects manipulated in the video material. The teaching-quality aspects were chosen on the basis of the availability of the five classroom videos used in this study and the teaching content presented in them. However, these choices limited the significance of our findings to the teaching-quality aspects examined and do not provide a comprehensive examination of all dimensions of teaching quality. In the future, assessments of other teaching-

quality aspects should be included in research on 360-degree classroom videos. Furthermore, the amount of interaction between teachers and students should be experimentally manipulated to investigate our explanations of the differences we found.

**Implications and Future Research**

With the present study, we aimed to contribute to a deeper understanding of how observers experience and assess teaching in different video environments. This knowledge can enable teacher educators and researchers to design classroom video environments so that teacher training and the assessment of teaching quality through classroom observation works best. Our findings support the assumption that immersive video environments offer a highly motivating and more realistic classroom experience (higher cognitive, affective, and physiological involvement) where observers put increased effort into their observations. For this reason, immersive video environments in teacher education could profitably augment traditional video-based learning environments. As competencies are situation-specific (Klieme et al., 2007), the realistic classroom experience in 360-degree VR videos might allow teachers' competencies to be trained in settings that are closer to the actual classroom and might therefore be particularly suitable for preparing PSTs for the complexity of real classrooms. Furthermore, our results hint at the special role of immersive video environments for observing classroom interactions between teachers and students. The use of 360-degree videos in teaching-quality research could be significant for examining interactions in the classroom, which is where teaching quality manifests at its core. If teachers see classroom interactions "better" in 360-degree videos, this result would also have an impact on the validation of important theories in the field (e.g., on teaching quality). For example, factor structures could emerge more clearly or effects in intervention studies could be better identified.

Future research on 360-degree classroom videos should distinguish the effects of video type and device by comparing multiple conditions that combine different types of videos (traditional and 360-degree videos) observed via different devices (standard PC and VR glasses) with large sample sizes. Furthermore, research on 360-degree classroom videos should supplement previous findings, which have largely been based on self-report data, with more objective measures (e.g., physiological measures). The first study to incorporate physiological data in the context of immersive classroom video environments was Ferdig et al.'s (2023a) study, which used Fitbits and heart rate variance (HRVa). Furthermore, future research should explore the role of ambisonic audio for classroom observations (Ferdig et al., 2023b). As Rupp et al. (2019) described 360-degree videos as less immersive than real VR, future research should

compare immersive 360-degree videos and real VR. Real VR is characterized by simulated virtual environments where users have agency (e.g., Makransky & Peterson, 2021). Simulated VR classrooms have already been applied in teacher education and research (e.g., Hasenbein et al., 2022; Richter et al., 2022). Exploring the conditions in which each type of immersive classroom experience is better suited for teacher training would facilitate the targeted use of the various innovative technologies.

**References**

Ainley, J., & Carstens, R. (2018). *Teaching and Learning International Survey (TALIS) 2018 Conceptual Framework.* OECD Education Working Papers Series, No. 187, OECD Publishing, Paris. http://dx.doi.org/10.1787/799337c2-en

Ardisara, A., & Fung, F. M. (2018). Integrating 360° videos in an undergraduate chemistry laboratory course. *Journal of Chemical Education*, *95*, 1881–1884. https://pubs.acs.org/doi/10.1021/acs.jchemed.8b00143

Atal, D., Admiraal, W., & Saab, N. (2023). 360° Video in teacher education: A systematic review of why and how it is used in teacher education. *Teaching and Teacher Education*, *135*, 104349. https://doi.org/10.1016/j.tate.2023.104349

Balzaretti, N., Ciani, A., Cutting, C., O'Keeffe, L., & White, B. (2019). Unpacking the potential of 360degree video to support pre-service teacher development. *Research on Education and Media*, *11*(1), 63–69. https://doi.org/10.2478/rem-2019-0009

Breves, P., & Schramm, H. (2021). Bridging psychological distance: The impact of immersive media on distant and proximal environmental issues. *Computers in Human Behavior*, *115*, 106606. https://doi.org/10.1016/j.chb.2020.106606

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, *2*(1), 49–68. https://doi.org/10.1177/109442819921004

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (2020). Between authenticity and cognitive demand: Finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teaching and Teacher Education*, *95*, 103–146. https://doi.org/10.1016/j.tate.2020.103146

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hoboken: Taylor and Francis.

Cortina, K. S., Müller, K., Häusler, J., Stürmer, K., Seidel, T., & Miller, K. F. (2018). Feedback mit eigenen Augen: Mobiles Eyetracking in der Lehrerinnen- und Lehrerbildung [Feedback with your own eyes: Mobile eye tracking in teacher education]. *Beiträge zur Lehrerinnen- und Lehrerbildung, 36*(2), 208–222. https://doi.org/10.25656/01:17097

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who Sees What?: Conceptual Considerations on the Measurement of Teaching Quality from Different Perspectives. *Zeitschrift für Pädagogik Beiheft*, *1*, 138–155. https://doi.org/10.3262/ZPB2001138

Fauth, B., Herbein, E., & Maier, J. L. (2022). *Beobachtungsmanual zum Unterrichtsfeedbackbogen Tiefenstrukturen (2. aktualisierte Version) [Observation manual for the classroom feedback form deep structures (2. updated version)].* Institut für Bildungsanalysen Baden-Württemberg.

Ferdig, R. E., & Kosko, K. W. (2020). Implementing 360 video to increase immersion, perceptual capacity, and noticing. *TechTrends*. Epub ahead of print 10 June 2020. https://doi.org/10.1007/s11528-020-00522-3

Ferdig, R., Kosko, K. W., & Gandolfi, E. (2023a). Using Fitbits and heart rate variance (HRVa) to understand pre-service teacher experiences in extended reality. In E. Langran, P. Christensen, & J. Sanson (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 1173–1179). New Orleans, LA, United States: Association for the Advancement of Computing in Education (AACE). Retrieved August 17, 2023 from https://www.learntechlib.org/primary/p/221982/

Ferdig, R. E., Kosko, K. W., & Gandolfi, E. (2023b). Exploring the relationships between teacher noticing, ambisonic audio, and variance in focus when viewing 360 video. *Educational Technology Research and Development*, *71*, 881–899. https://doi.org/10.1007/s11423-023-10215-2

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., & Andersen, S. A. W. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial. *Surgical Endoscopy*, *34*, 1244–1252. https://doi.org/10.1007/s00464-019-06887-8

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

Gold, B., & Windscheid, J. (2020). Observing 360-degree classroom videos – Effects of video type on presence, emotions, workload, classroom observations, and ratings of teaching quality. *Computers & Education, 156*, 103960. https://doi.org/10.1016/j.compedu.2020.103960

Grossman, P. (2021). *Teaching core practices in teacher education.* Harvard Education Press.

Hajahmadi, S., & Marfia, G. (2023). Effects of the uncertainty of interpersonal communications on behavioral responses of the participants in an immersive virtual reality experience: A usability study. *Sensors*, *23*(4), 2148. https://doi.org/10.3390/s23042148

Hamel, C., & Viau-Guay, A. (2019). Using video to support teachers' reflective practice: A literature review. *Cogent Education*, *6*(1), 1673689. https://doi.org/10.1080/2331186X.2019.1673689

Harackiewicz, J. M., Smith, J. L., & Priniski, S. J. (2016). Interest matters: The importance of promoting interest in education. *Policy Insights From the Behavioral and Brain Sciences*, *3*(2), 220–227. https://doi.org/10.1177/2372732216655542

Hasenbein, L., Stark, P., Trautwein, U., Queiroz, A. C. M., Bailenson, J., Hahn, J. U., & Göllner, R. (2022). Learning with simulated virtual classmates: Effects of social-related configurations on students' visual attention and learning experiences in an immersive virtual reality classroom. *Computers in Human Behavior*, *133*, 107282. https://doi.org/10.1016/j.chb.2022.107282

Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *The Elementary School Journal*, *101*(1), 3–20.

Hofman, J. (2022). Classroom management and teacher emotions in secondary mathematics teaching: a qualitative video-based single case study. *Education Inquiry*, 1–17. https://doi.org/10.1080/20004508.2022.2028441

Huang, W. (2020). *Investigating the Novelty Effect in Virtual Reality on Stem Learning.* Doctoral dissertation, Arizona State University.

Huang, W., Roscoe, R. D., Johnson-Glenberg, M. C., & Craig, S. D. (2021). Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. *Journal of Computer Assisted Learning*, *37*(3), 745–758. https://doi.org/10.1111/jcal.12520

IBM Corp. (2020). IBM SPSS Statistics for Windows (Version 27.0) [Computer software]. IBM Corp.

Jacobs, J. K., Garnier, H., Gallimore, R., Hollingsworth, H., Givvin, K. B., Rust, K., & Stigler, J. W. (2003). *Third International Mathematics and Science Study 1999 Video Study Technical Report, Volume 1: Mathematics.* (NCES 2003012). Washington, D.C.: National Center for Education Statistics

Kilburn, D. (2014). *Methods for recording video in the classroom: producing single and multi-camera videos for research into teaching and learning.* (NCRM Working Paper). NCRM. Retrieved July 28, 2023 from http://eprints.ncrm.ac.uk/3599/

Klette, K. (2023). Classroom observation as a means of understanding teaching quality: towards a shared language of teaching? *Journal of Curriculum Studies*, 55(1), 49–62. https://doi.org/10.1080/00220272.2023.2172360

Klieme, E., Hartig, J., & Rauch, D. P. (2007). The concept of competence in educational contexts. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts* (pp. 3–22). Hogrefe & Huber Publishers.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (*pp*. 137–160). Waxmann.

Koch, M., Luck, K. V., Schwarzer, J., & Draheim, S. (2018). The novelty effect in large display deployments - experiences and lessons-learned for evaluating prototypes. Paper presented at: *Proceedings of 16th European Conference on Computer-Supported Cooperative Work-Exploratory Papers. European Society for Socially Embedded Technologies (EUSSET).* https://doi.org/10.18420/ecscw2018_3

Kosko, K. W., Ferdig, R. E., & Zolfaghari, M. (2021). Pre-service teachers' professional noticing when viewing standard and 360 video. *Journal of Teacher Education, 72*(3), 284–297. https://doi.org/10.1177/0022487120939544

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

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, *4*(1), 1280256. https://doi.org/10.1080/2331186X.2017.1280256

Kunz, K., & Zinn, B. (2022). Virtuelle Unterrichtsszenarien in der Lehrpersonenbildung - eine Studie zur Akzeptanz, Immersion und zum Präsenzerleben mit Studierenden der Berufs- und Technikpädagogik [Virtual teaching scenarios in teacher education - a study of acceptance, immersion, and classroom experience with vocational and technical education students]. *Unterrichtswissenschaft*, 1–25. https://doi.org/10.1007/s42010-022-00151-0

Liu, R., Wang, L., Koszalka, T. A., & Wan, K. (2022). Effects of immersive virtual reality classrooms on students' academic achievement, motivation and cognitive load in science lessons. *Journal of Computer Assisted Learning*, *38*(5), 1422–1433. https://doi.org/10.1111/jcal.12688

Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development*, *2010*(5), 1–24. https://doi.org/10.1007/s11423-018-9581-2

Makransky, G., & Petersen, G. B. (2021). The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, 1–22. https://doi.org/10.1007/s10648-020-09586-2

Meyer, O. A., Omdahl, M. K., & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education*, *103603*, 103603. https://doi.org/10.1016/J.COMPEDU.2019.103603

Paulicke, P., Ehmke, T., Pietsch, M., & Schmidt, T. (2019). Wie beeinflusst die Kameraperspektive die Beurteilung der Unterrichtsqualität? [How does the camera perspective influence the assessment of teaching quality?]. *Zeitschrift für Bildungsforschung*, *9*(3), 411–435. https://doi.org/10.1007/s35834-019-00246-2

Peterson, S. M., Furuichi, E., & Ferris, D. P. (2018). Effects of virtual reality high heights exposure during beam-walking on physiological stress and cognitive loading. *PloS one*, *13*(7), e0200306. https://doi.org/10.1371/journal.pone.0200306

Piwowar, V., Barth, V. L., Ophardt, D., & Thiel, F. (2018). 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., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Prilop, C. N., Weber, K. E., & Kleinknecht, M. (2020). Effects of digital video-based feedback environments on pre-service teachers' feedback competence. *Computers in Human Behavior*, *102*, 120–131. https://doi.org/10.1016/j.chb.2019.08.011

Ranieri, M., Luzzi, D., Cuomo, S., & Bruni, I. (2022). If and how do 360 videos fit into education settings? Results from a scoping review of empirical research. *Journal of Computer Assisted Learning*, *38*(5), 1199–1219. https://doi.org/10.1111/jcal.12683

Richter, E., Hußner, I., Huang, Y., Richter, D., & Lazarides, R. (2022). Video-based reflection in teacher education: Comparing virtual reality and real classroom videos. *Computers & Education*, *190*, 104601. https://doi.org/10.1016/j.compedu.2022.104601

Rigby, J. M., Brumby, D. P., Gould, S. J., & Cox, A. L. (2019). Development of a questionnaire to measure immersion in video media: The Film IEQ. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video* (pp. 35–46). https://doi.org/10.1145/3317697.3323361

Roche, L., Cunningham, I., & Rolland, C. (2023). 360° Video uses in teacher education: A literature review. In E. Langran, P. Christensen & J. Sanson (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 1205–1213). New Orleans, LA, United States: Association for the Advancement of Computing in Education (AACE). Retrieved December 1, 2023 from https://www.learntechlib.org/primary/p/221987/

Rupp, M. A., Odette, K. L., Kozachuk, J., Michaelis, J. R., Smither, J. A., & McConnell, D. S. (2019). Investigating learning outcomes and subjective experiences in 360-degree videos. *Computers & Education*, *128*, 256–268. https://doi.org/10.1016/j.compedu.2018.09.015

Ruth-Herbein, E., Maier, J. L., Fauth, B. (2022). Promoting Teaching Quality Through Classroom Observation and Feedback: Design of a Program in the German State of Baden-Württemberg. In J. Manzi, Y. Sun, & M. R. García (Eds.) *Teacher Evaluation Around the World. Teacher Education, Learning Innovation and Accountability.* (pp. 271–289). Springer, Cham. https://doi.org/10.1007/978-3-031-13639-9_12

Santagata, R., & Guarino, J. (2011). Using video to teach future teachers to learn from teaching. *ZDM, 43*, 133–145. https://doi.org/10.1007/s11858-010-0292-3

Seidel, T., Farrell, M., Martin, M., Rieß, W., & Renkl, A. (2022). Developing scripted video cases for teacher education: Creating evidence-based practice representations using mock ups. *Frontiers in Education*, *7*, 965498. http://dx.doi.org/10.3389/feduc.2022.965498

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

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Snelson, C., & Hsu, Y. C. (2020). Educational 360-degree videos in virtual reality: A scoping review of the emerging research. *TechTrends*, *64*(3), 404–412. https://doi.org/10.1007/s11528-019-00474-3

Stevens, J. P. (2012). Applied multivariate statistics for the social sciences (5th ed.). New York, NY: Taylor and Francis Group.

Walshe, N., & Driver, P. (2019). Developing reflective trainee teacher practice with 360-degree video. *Teaching and Teacher Education*, 78, 97–105. https://doi.org/10.1016/j.tate.2018.11.009

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

Windscheid, J., & Will, A. (2018). *A web-based multi-screen 360-degree video player for pre-service teacher training.* Universitätsbibliothek. http://dx.doi.org/10.6084/m9.figshare.6527309.v1

Wirth, W., Schramm, H., Böcking, S., Gysbers, A., Hartmann, T., Klimmt, C., & Vorderer, P. (2008). Entwicklung und Validierung eines Fragebogens zur Entstehung von Räumlichem Präsenzerleben [Development and validation of a questionnaire on the emergence of spatial presence experience]. In J. Matthes, W. Wirth, G. Daschmann, & A. Fahr (Eds.), *Die Brücke zwischen Theorie und Empirie: Operationalisierung, Messung und Validierung in der Kommunikationswissenschaft* (pp. 70–95). Halem.

Zachrich, L., Weller, A., Baron, C., & Bertram, C. (2020). Historical experiences: A framework for encountering complex historical sources. *History Education Research Journal, 17*(2), 243–275. https://doi.org/10.14324/HERJ.17.2.08

Zee, M., & Koomen, H. M. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational Research*, *86*(4), 981–1015. https://doi.org/10.3102/0034654315626801

## Appendix

**Table A1**

*Univariate Main Effects of Video Environment for Individual Outcome Variables*

| Within-subjects effect | Outcome | Type III sum of squares | *df* | Mean square | *F* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|---|
| Video environment | Immersion | 37.95 | 1 | 37.95 | 56.34 | < .001 | .44 |
| | Focus of attention | 1.83 | 1 | 1.83 | 5.53 | .021 | .07 |
| | Compassion for teacher | 0.51 | 1 | 0.51 | 0.60 | .442 | .01 |
| | Compassion for students | 13.40 | 1 | 13.40 | 17.79 | < .001 | .20 |
| | Self-location | 226.49 | 1 | 226.49 | 478.82 | < .001 | .87 |
| | Possible actions | 94.50 | 1 | 94.50 | 127.44 | < .001 | .64 |
| | Mental load | 1.22 | 1 | 1.22 | 0.73 | .395 | .01 |
| | Mental effort | 3.75 | 1 | 3.75 | 11.13 | .001 | .13 |

**Table A2**

*Univariate Main Effects of Novelty and Effects of Novelty × Video Environment Interaction for Individual Outcome Variables*

| | Within-subjects | Measure | Type III sum of squares | *df* | Mean square | *F* | *p* | $\eta^2$ |
|---|---|---|---|---|---|---|---|---|
| Between-subjects effect | Novelty | Immersion | 0.31 | 1 | 0.31 | 0.30 | .588 | .00 |
| | | Focus of attention | 1.83 | 1 | 1.83 | 2.74 | .102 | .04 |
| | | Compassion for teacher | 0.02 | 1 | 0.02 | 0.01 | .914 | .00 |
| | | Compassion for students | 6.01 | 1 | 6.01 | 6.02 | .017 | .08 |
| | | Self-location | 4.65 | 1 | 4.65 | 5.49 | .022 | .07 |
| | | Possible actions | 3.59 | 1 | 3.59 | 3.18 | .079 | .04 |
| | | Mental load | 0.20 | 1 | 0.20 | 0.12 | .73 | .00 |
| | | Mental effort | 0.13 | 1 | 0.13 | 0.10 | .76 | .00 |
| Within-subjects effect | Video Environment ×Novelty | Immersion | 3.30 | 1 | 3.30 | 4.90 | .030 | .06 |
| | | Focus of attention | 1.31 | 1 | 1.31 | 3.96 | .050 | .05 |
| | | Compassion for teacher | 0.51 | 1 | 0.51 | 0.60 | .442 | .01 |
| | | Compassion for students | 1.03 | 1 | 1.03 | 1.36 | .247 | .02 |
| | | Self-location | 0.00 | 1 | 0.00 | 0.00 | .988 | .00 |
| | | Possible actions | 2.60 | 1 | 2.60 | 3.51 | .065 | .05 |
| | | Mental load | 2.05 | 1 | 2.05 | 1.23 | .271 | .02 |
| | | Mental effort | 0.06 | 1 | 0.06 | 0.18 | .676 | .00 |

**Table A3**

*Univariate Main Effects of Novelty and Effects of Novelty × Video Environment Interaction for Individual Teaching-Quality Aspects*

| | Focus on key concepts | | | Challenging tasks | | | Classroom disruptions | | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimate | SE | *p* | Estimate | SE | *p* | Estimate | SE | *p* |
| Novelty | .01 | .05 | .756 | -.02 | .09 | .859 | .00 | .04 | .923 |
| Novelty×Video Environment | -.02 | .05 | .758 | -.04 | .11 | .731 | .06 | .06 | .295 |

# 5

# STUDY 3

Daltoè, T., Appel, T., Stark, P., Brucker, B., Dreher, A., Fauth, B., Friesen, M., Gerjets, P., Hansen, L., Trautwein, U., & Göllner, R. (2024). *Connecting gaze behavior and ratings of teaching quality*. Manuscript submitted for publication.

**Abstract**

Teaching quality is commonly assessed through external raters' observations of classroom videos. However, these ratings frequently exhibit limited psychometric quality. Beyond evaluating the ratings themselves, exploring the observation process and the design of the video environment may offer valuable insights into the conditions that enhance rating accuracy. With this study, we aimed to explore the classroom observation process using eye-tracking technology and to link observers' gaze behavior with the accuracy of their teaching-quality ratings. Additionally, we examined the impact of different video environments by comparing traditional classroom videos presented on computer screens with immersive 360-degree classroom videos presented on virtual-reality headsets. $N = 75$ pre-service teachers participated in a controlled lab study. Each participant observed two randomly assigned classroom videos—one in a traditional and one in an immersive video environment. Eye trackers recorded gaze behavior during critical classroom events. Participants rated the quality of the observed teaching after the video observations. Our results indicate that observers distributed their visual attention differently, depending on the type of critical classroom event. Observers' visual focus of attention in the videos as well as cognitive engagement during critical events, indicated by a larger pupil diameter, partly predicted the accuracy of teaching-quality ratings. The associations between gaze behavior and rating accuracy were stronger in immersive 360-degree classroom videos. Observers' cognitive processes during classroom observations hold great potential for understanding their assessments of teaching quality.

## Introduction

In educational research, the use of classroom videos is the gold standard for investigating teaching and learning in the classroom (e.g., Janik & Seidel, 2009). In video-based classroom research, observers usually watch a classroom video and draw conclusions about the observed teaching, depending on the aim of the specific classroom observation. In many contexts, observers provide ratings of teaching quality using standardized classroom observation systems (Bell et al., 2019; Gitomer, 2021). Although observer ratings of teaching quality are a well-established methodology, significant challenges with the psychometric quality of these ratings have been identified. Empirical studies have revealed low reliabilities and rater bias (e.g., halo effects) in these ratings (Praetorius et al., 2012), likely resulting in less power for predicting teaching-relevant outcomes (e.g., student learning outcomes). Thus, further research is necessary to pinpoint the factors that explain these inconsistencies in observer ratings of teaching quality.

In this article, we aim to gain a deeper understanding of observer ratings of teaching quality by breaking down the process of teaching-quality assessments in classroom observations. First, we describe the relevant steps observers must follow to rate teaching quality in video-based classroom observations. We assume that an accurate assessment of teaching quality depends on observers noticing the critical events in the classroom (Santagata et al., 2021) and drawing the right conclusions about teaching quality from these events. Thus, we used observers' gaze behavior during classroom observations to investigate the process of classroom observation and connected this gaze behavior to the accuracy of observer ratings. Second, we assume that the complexity of the information provided in the observation environment is crucial for the assessment of teaching quality (Atal et al., 2023). For this reason, we varied the video environment (observation of traditional videos on a standard desktop PC vs. 360-degree videos with virtual reality [VR] headsets) to examine the accuracy of observers' assessments of teaching quality. With this study, our goal is to enhance the current body of literature on observer ratings of teaching quality by gaining deeper insights into observers' cognitions during classroom observations in different video environments.

### Observer Ratings of Teaching Quality

International video studies such as TIMSS (Hiebert & Stigler, 2000), or the German-Swiss Pythagoras study (Klieme et al., 2009) have yielded compelling findings on teaching effectiveness.

For example, video studies have highlighted the significance of basic dimensions of teaching quality (Klieme et al., 2009; Praetorius et al., 2018), such as cognitive activation and classroom management, for student learning outcomes (Fauth et al., 2014; Lipowsky et al., 2009). In video studies, trained raters observe classroom videos, analyze the teaching, and quantify their observations using standardized classroom observation systems (Bell et al., 2019). Examples of widely adopted observation systems include the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008), or the Protocol for Language Arts Teaching Observations (PLATO; Grossman et al., 2013).

Although observer ratings of teaching quality are widely used, research has highlighted significant challenges in their psychometric quality. Studies continue to document substantial disagreements among raters and low to moderate reliabilities for ratings (Bell et al., 2019). Different arguments have been used to explain the limited psychometric quality of observer ratings. On the one hand, the level of training raters received for classroom observations (Bergin et al., 2017) and the degree of alignment between raters' expertise and the observed teaching subject (Dreher & Leuders, 2021; Schlesinger et al., 2018) have been considered. On the other hand, factors such as the observation setting (in-classroom vs. video-based; Jentsch et al., 2024) and the camera angle used in video-based observations (Paulicke et al., 2019) have also contributed to the inconsistent psychometric quality. In addition to these existing approaches, it may be beneficial to examine the process by which raters observe teaching in video-based classroom observations and how this observation process is linked to ratings of teaching quality.

**Process of Classroom Observation**

Analyzing teaching in classroom observations and providing ratings of teaching quality is a complex process (Bell et al., 2012). Whereas research has explored the psychometric quality of ratings (e.g., Praetorius et al., 2012), it is equally important to investigate the classroom observation process itself to understand the steps observers must follow to make their ratings.

An accurate assessment of teaching quality in classroom observations relies on successfully noticing and correctly interpreting classroom events that determine teaching quality (Santagata et al., 2021). A prerequisite for noticing critical events is being able to visually identify them in the first place (Just & Carpenter, 1976). For this reason, the first step of classroom observation consists of observers directing their visual attention toward relevant activities and interactions in the

classroom. Relevant information can arise from either the teacher or the students, making it essential to allocate attention effectively to the appropriate locations. In research on teachers' professional vision, gaze fixations on specific areas of interest (AOIs) serve as an indicator of teachers' visual attention to the critical events in the classroom (McIntyre et al., 2019; Stürmer et al., 2017) or in classroom videos (Grub et al., 2022; Stahnke & Blömeke, 2021). Additionally, gaze transitions between AOIs provide insights into teachers' visual attention control in the classroom (Kosel et al., 2021). Using observers' gaze as an indicator of visual attention on critical classroom events is also applicable for investigating the process of classroom observation.

After directing their attention to relevant activities and interactions, observers must identify them as significantly affecting students' learning experiences and, consequently, as essential information for evaluating teaching quality. A pertinent measure from eye-tracking technology that might be associated with the identification of critical events is the pupil diameter, which is associated with cognitive load (Appel et al., 2018; Jainta & Baccino, 2010), information processing (Goldinger & Papesh, 2012), and emotional arousal (Wang et al., 2018). In the context of classroom observations, a larger pupil diameter may suggest that observers are more deeply engaged both cognitively and emotionally with the observed event. This engagement might affect the inferences they draw from that event in their teaching-quality ratings.

To quantify teaching quality in classroom observations, observers usually provide ratings via a standardized classroom observation system with an underlying theoretical framework of teaching quality (Bell et al., 2019). Observation systems comprise several theoretically distinct aspects of teaching quality that observers are asked to rate on a numeric scale. Therefore, after detecting critical events in the teaching, observers need to categorize the detected events in terms of the theoretically distinct teaching-quality aspects in the respective classroom observation system. After determining which aspect of teaching quality the observed events pertain to, observers must assess the quality of these events and assign a score on a numeric scale (e.g., 1–4). This process requires thorough knowledge of the observation system's scoring procedures (Bell et al., 2019).

Considering the steps observers must take to assess teaching quality in classroom observations, it is clear that gaze-related measures (e.g., visual attention to and transitions between

specific AOIs in the classroom; pupil diameter) could provide valuable insights into the observation process and clarify how observers arrive at their assessments of teaching quality.

**Classroom Observation in Different Video Environments**

All of the steps involved in the process of classroom observation occur in an environment of complex information. Whereas observations in the actual classroom potentially provide all the information needed to achieve an accurate rating, video-based observations allow for a more focused and analytic perspective on the classroom (Janik & Seidel, 2009). However, the range of information available for observers to base their ratings on may be limited in video-based assessments.

Recent developments in video technology have made it possible to use different kinds of videos that come with varying amounts of visual information. Traditionally, classroom videos have been presented on standard PCs using 4:3 or 16:9 video formats. More recently, initial studies explored 360-degree classroom videos as an innovative video environment for classroom observations (Atal et al., 2023). Such videos capture the whole classroom and allow video observations without a fixed camera angle (restricted field of view). They promise a more immersive classroom observation experience with VR headsets (Atal et al., 2023).

The first studies on 360-degree classroom videos suggest that immersive 360-degree videos might help observers notice specific teaching content (Kosko et al., 2021). However, the detection of relevant classroom events depends on the field of view observers choose in the 360-degree video (Kosko et al., 2022). These research findings suggest both the potential and pitfalls of 360-degree classroom videos for observer ratings of teaching quality. On the one hand, helping observers notice critical events might help generate accurate assessments of teaching quality. On the other hand, it might be easier to miss critical events when they are not in the observers' field of view, thus potentially leading to less accurate assessments of teaching quality.

Initial studies comparing observer ratings of teaching quality in different video environments found no or small rating differences between traditional and immersive video environments (Daltoè et al., 2024; Gold & Windscheid, 2020). However, previous studies on 360-degree classroom videos did not utilize eye-tracking technology to explore how gaze is related to noticing events or assessing teaching quality. This point is particularly relevant in 360-degree

video environments (due to the self-regulated choice of viewing direction), but it also applies to traditional videos where viewers can also look at different areas or even look away. Therefore, by analyzing gaze behavior and its relationship to rating quality in different video environments, this study substantially extends the current body of literature.

**Present Study and Research Questions**

In this study, we investigated how observers arrive at their assessments of teaching quality in video-based classroom observations. We assumed that noticing and interpreting critical events during classroom observation is a prerequisite for accurate ratings of teaching quality. In Research Question 1 (RQ1), we looked at observers' gaze behavior during the observation of critical events to investigate whether observers distributed their visual attention in a way that enabled them to detect these events in the first place. We classified the critical events by whether the teacher or the students conducted the event. We asked: Do observers show different gaze behavior for critical classroom events conducted by the teacher versus the students (RQ1)?

We hypothesized that different events would result in different gaze behavior. We expected that observers would visually focus more on the teacher during events conducted by the teacher and would focus more on the students during events conducted by the students.

In RQ2, we aimed to understand the connection between observers' gaze behavior and the accuracy of their teaching-quality ratings. We asked whether observers' gaze-related indicators (visual attention on teacher- and student-centered AOIs, transitions between AOIs, pupil diameter) predict accurate teaching-quality ratings: Is observers' gaze behavior associated with the accuracy of resulting teaching-quality ratings (RQ2)?

Regarding visual attention, we hypothesized that observers focusing on the appropriate areas in the classroom would be able to provide more accurate teaching-quality ratings. For more teacher-focused aspects of teaching quality, we expected that a stronger focus on the teacher-centered AOI would result in more accurate ratings. For more student-focused aspects of teaching quality, we expected that a stronger focus on the student-centered AOI would result in more accurate ratings. Regarding pupillometry, we hypothesized that observers with a greater pupil diameter—indicating higher levels of information processing—would provide more accurate teaching-quality ratings.

In RQ3, we included the complexity of information provided by the video environment as a factor potentially affecting classroom observations and teaching-quality ratings. We compared the associations investigated in RQ1 and RQ2 between a traditional and an immersive 360-degree video environment: Are there differences in associations with gaze behavior (RQ1) and associations with rating accuracy (RQ2) between traditional and immersive video environments (RQ3)?

We had no directed hypotheses for effects regarding the video environments.

## Method

This study was conducted during the winter term (December 2022 to February 2023) at the Leibniz-Institut für Wissensmedien in Tübingen. The study design was preregistered on aspredicted.org (https://aspredicted.org/67J_93Q). With the same data, we generated a previous paper (Daltoè et al., 2024) focusing on comparing self-reported classroom observation experiences and teaching-quality ratings between traditional and immersive video environments. The present paper utilizes eye-tracking technology to explore observers' gaze behavior during classroom observations and how it is related to the accuracy of observer ratings of teaching quality.

### Participants

The observers in this study were $N = 75$ pre-service teachers (72% female, 28% male; $M_{age} = 24.72$, $SD = 2.02$) from two universities who already had teaching experience. They were preparing to teach different subjects (e.g., $n = 13$ German language arts, $n = 10$ mathematics, $n = 5$ physics). Recruitment was conducted through university mailing lists, subject-specific mailing lists, and advertisements in courses for pre-service teachers in master's programs. Participants were paid 20€, a common practice in similar studies. Our sample size was chosen to be comparable to similar studies in the field.

### Procedure

Participants provided informed consent upon arrival at the lab and were fitted with wristbands for physiological data collection. They were then introduced to the classroom observation task through a video lasting 8:46 min. Subsequently, each participant observed two classroom videos: one in an immersive 360-degree video environment using a VR headset and one

in a traditional video environment on a standard PC. The allocation of video content and the order of video environments were randomized for each participant. In the immersive video environment, observers were presented 360-degree videos from a viewpoint in the center of the classroom. From this viewpoint, observations in all directions were possible by moving the head. In the traditional video environment, observers were presented split screen videos, showing the teaching from two camera angles to capture the whole classroom. Following each observation, participants completed questionnaires assessing their observation experience and teaching-quality ratings. The experimental session lasted approximately 90 min, including questionnaires.

**Apparatus**

In the traditional video environment, we used laptops and recorded eye-tracking data using SMI Red remote eye trackers with a sampling frequency of 250 Hz. The eye-tracker was calibrated using a 9-point calibration before the video observation. In the immersive video environment, we used HP Reverb G2 Omnicept Edition VR headsets with integrated Tobii eye trackers with a sampling frequency of 120 Hz. The eye-tracker was also calibrated using a 9-point calibration before the video observation. In each video environment, lighting conditions in the lab were kept constant throughout the observation.

**Video Material and Critical Events**

Participants observed two out of a total of five staged classroom videos, each lasting between 6:33 and 9:06 min, from two German mathematics classrooms. These videos were developed by a team of researchers and practitioners, including experts in mathematics education and teaching-quality research. The five staged teaching situations included 30 critical events (ranging from three to nine events per video) considered essential for accurately rating teaching quality. By using these staged videos with predefined critical events, we provided targeted evidence for assessing three aspects of teaching quality: two more teacher-focused aspects of cognitive activation: *focus on key concepts* and *challenging tasks*; and one more student-focused aspect of classroom management: *classroom disruptions*. These three aspects of teaching quality are theoretically grounded in Fauth et al.'s (2022) classroom observation system, which was used to evaluate teaching quality in this study. This alignment between the presented critical events and the theoretical framework of teaching quality enabled us to deliberately link the observed critical events to the teaching-quality ratings.

The critical events in the staged videos can be classified on the basis of their focus on either the students' or the teacher's behavior and the aspect of teaching quality they exemplified. Based on this classification, a different visual focus of attention (students vs. teacher) is necessary to notice the events and assess teaching quality accurately. Table 1 presents three examples of event classifications. Table A1 in Appendix A presents a complete list of the critical events and their classifications.

**Table 1**

*Examples of Classifications of Critical Events*

| Critical event examples | Critical event enacted by teacher or students | Critical event exemplifies this aspect of teaching quality |
|---|---|---|
| The teacher draws apples and pears on the board. | Teacher | Focus on key concepts |
| The teacher asks: "Can we derive a general formula from these ideas to calculate the area of any given circle sector? Take a look at the examples on the board again." | Teacher | Challenging tasks |
| Students throw paper balls while the teacher is drawing a line segment on the board. | Students | Classroom disruptions |

## Measures

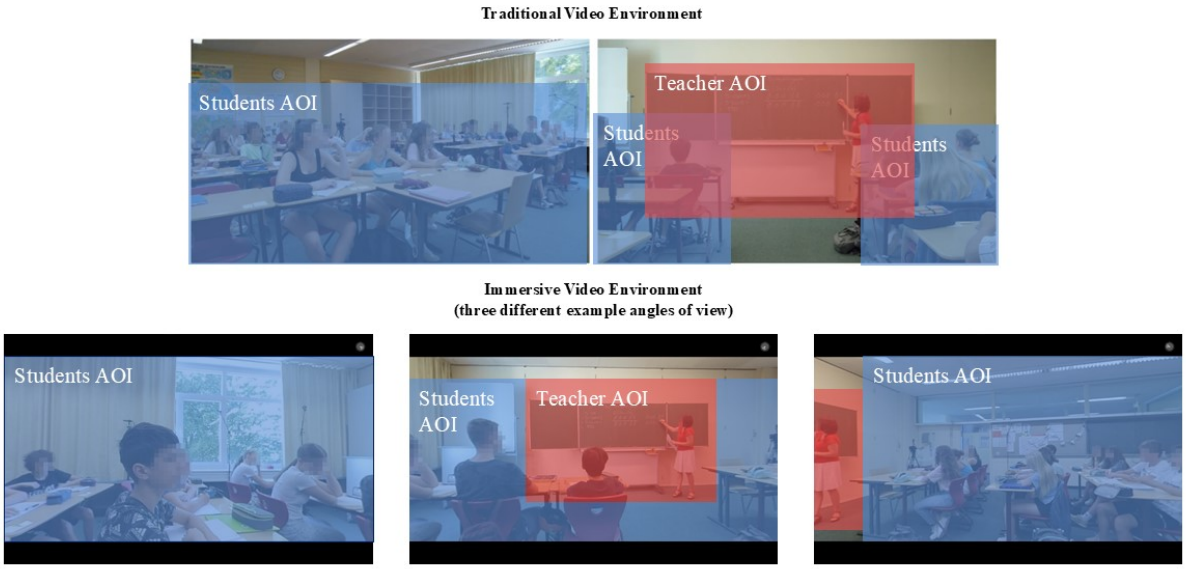
### *Eye-Tracking Measures*

We used comparable measures from the eye-tracking data in both traditional and immersive video environments. Specifically, we measured overt visual attention directed toward the teacher and students with two AOIs: teacher-centered and student-centered AOIs in both video conditions. Figure 1 shows an example of how the AOIs were defined. To capture all the teacher's actions,

sometimes the students also fell into the teacher-centered AOI due to the position of the 360-degree camera. We made sure to keep the AOIs analogous in the two video environments (see Figure 1). To assess visual attention toward the teacher and students, we calculated the dwell time on the AOIs during quality-relevant teaching events. Because the events varied in duration, the relative proportion of dwell time on each AOI per event was based on the event duration. To determine changes in visual attention between the AOIs during events, we counted the number of transitions between the AOIs. Additionally, we calculated the observers' event-specific pupil diameter in millimeters.

**Figure 1**

*Examples of Teacher- and Student-Centered AOIs in Both Video Environments*




### *Observer Ratings of Teaching Quality*

Observer ratings of teaching quality were based on the standardized classroom observation system by Fauth et al. (2022). They focused on 11 teaching-quality items grounded in the framework of the basic dimensions of teaching quality: cognitive activation, student support, and classroom management (Klieme et al., 2009; Praetorius et al., 2018). The teaching-quality items were rated on a 4-point scale, ranging from 1 (*totally disagree*) to 4 (*totally agree*). Participants were also given the opportunity to provide open-ended explanations for their teaching-quality ratings. As the video material in this study focused on critical events that were related to three of

the 11 teaching-quality items from the observation system, we analyzed only the ratings for these three items: *focus on key concepts* (cognitive activation), *challenging tasks* (cognitive activation), and *classroom disruptions* (classroom management). As an indicator of rating accuracy, we calculated the absolute deviation of the individual rating from a master rating for the staged classroom videos. Master ratings were created collaboratively by the experts who developed the staged classroom videos.

**Statistical Analyses**

Data analyses were conducted with R version 4.4.0. Due to the multilevel structure of our data (events nested in videos, videos nested in participants), multilevel regression models were calculated. To address the question of whether different critical events resulted in different gaze behavior when observing these events (RQ1), the first set of analyses predicted the assessed eye-tracking measures (dwell time on teacher- and student-centered AOIs, transitions between AOIs, pupil diameter). Whether the event focused on student- or teacher-centered AOIs served as the independent variable, controlling for the video environment and raters' subject-specific background, as not all participants had backgrounds in mathematics or physics, which fit the teaching content.

To address the question of whether observers' gaze behavior was associated with the accuracy of teaching-quality ratings (RQ2), we predicted the deviation from a master rating for each of the three teaching-quality aspects. The focus of the event and all eye-tracking measures (proportion of dwell time on the teacher-/student-centered AOIs, transitions between AOIs, pupil diameter) served as independent variables, controlling for the video environment and the raters' subject-specific background.

To address the question of whether the differences in visual attention and associations between gaze behavior and rating accuracy differed between video environments (RQ3), we included interactions between the contrast-coded video environment variable (traditional vs. immersive) and event focus, eye-tracking measures, and subject-specific background in the regression models from RQ1 and RQ2 to conduct a statistical comparison of the results between video environments. One-tailed tests were used for directional hypotheses; two-tailed tests were used for nondirectional hypotheses. The alpha level was set to .05.

## Results

Table 2 presents the descriptive statistics for the eye-tracking measures and deviations from master ratings for the three aspects of teaching quality for the traditional and immersive video environments (PC and VR). Table B1 in Appendix B presents correlations for all metric variables.

**Table 2**

*Descriptive Statistics for the Eye-Tracking Measures and Rating Accuracies Presented Separately for the Traditional (PC) and Immersive (VR) Video Environments*

| | PC | | VR | |
|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* |
| Proportion of dwell time on teacher-centered AOI | 40.80 | 31.87 | 54.55 | 39.06 |
| Proportion of dwell time on student-centered AOI | 60.24 | 31.90 | 45.01 | 38.96 |
| Transitions between AOIs | 4.84 | 4.56 | 4.31 | 6.03 |
| Pupil diameter | 3.38 | 0.37 | 3.97 | 0.75 |
| DMR: Focus on key concepts | 0.76 | 1.02 | 0.57 | 0.84 |
| DMR: Challenging tasks | 0.91 | 0.97 | 0.82 | 0.93 |
| DMR: Classroom disruptions | 0.16 | 0.40 | 0.14 | 0.34 |

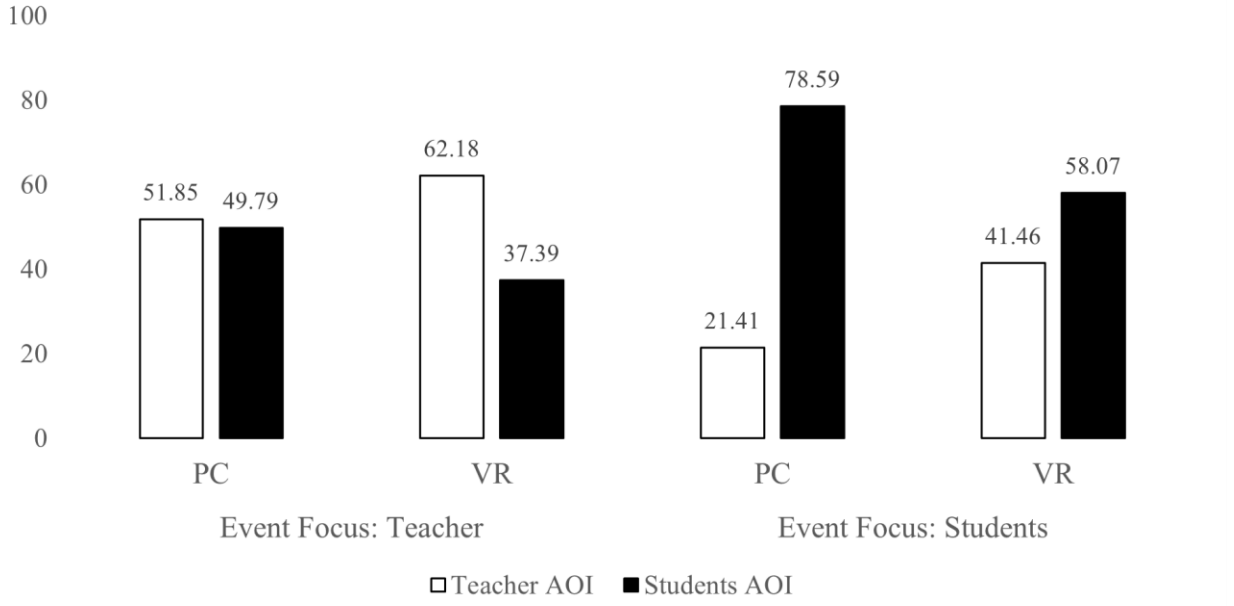
*Note.* DMR = deviation from master rating in scale points on a 4-point scale.

### Visual Attention Toward Relevant Classroom Events

In RQ1, we investigated whether observers' gaze behavior differed between critical classroom events conducted by the teacher versus the students. Figure 2 presents descriptive results for the proportion of dwell time on teacher- and student-centered AOIs by the focus of the event on the teacher and the students, separately for each video environment.

**Figure 2**

*Proportion of Dwell Time on Teacher and Students for Critical Events Focusing on the Teacher's and Students' Behavior in the Traditional (PC) and Immersive (VR) Video Environments*




The regression models revealed that observers differed significantly in their visual attention toward the teacher and students depending on the focus of the event (see Table 3). As expected, during events consisting of students' behavior, observers focused more on the students and less on the teacher and vice versa (ß = -0.69, *SE* = 0.07, *t* = -9.41, *p* < .001). Also, in the immersive video environment, observers focused significantly more on the teacher overall (ß = 0.17, *SE* = 0.04, *t* = 3.91, *p* < .001). We found no differences in transitions or pupil diameter for events conducted by the teacher or the students. Pupil diameter was larger overall in the immersive video environment (ß = 0.43, *SE* = 0.03, *t* = 16.08, *p* < .001), which can be explained by the difference in lighting conditions between video environments.

**Table 3**

*Predicting Observers' Gaze Behavior From Event Focus, Video Environment, and Subject-Specific Background*

| | Ratio of dwell time on teacher- and student-centered AOIs | | | | Transitions between AOIs | | | | Pupil diameter | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* |
| Intercept | 0.28 | 0.12 | 2.35 | .07 | -0.03 | 0.17 | -0.16 | .88 | -0.06 | 0.19 | -0.33 | .75 |
| Event focus[a] | **-0.69** | **0.07** | **-9.41** | **< .001*** | -0.12 | 0.08 | -1.48 | .14 | -0.02 | 0.04 | -0.50 | .62 |
| Video environment[b] | **0.17** | **0.04** | **3.91** | **< .001** | -0.09 | 0.05 | -1.94 | .05 | **0.43** | **0.03** | **16.08** | **< .001** |
| Subject-specific background[c] | -0.06 | 0.12 | -0.48 | .64 | 0.04 | 0.13 | 0.28 | .78 | 0.12 | 0.22 | 0.54 | .59 |

[a] 0 = teacher, 1 = students; [b] 0 = traditional, 1 = immersive; [c] 0 = no background in mathematics/physics, 1 = background in mathematics/physics.

*one-tailed testing based on a directional hypothesis.

**Associations Between Gaze Behavior and Rating Accuracy**

In RQ2, we investigated whether gaze behavior during classroom observations was associated with the accuracy of the resulting teaching-quality ratings. The regression models revealed significant associations between gaze behavior and rating accuracies for the two more teacher-focused aspects of teaching quality: *focus on key concepts* and *challenging tasks* (see Table 4). For the aspect *focus on key concepts*, pupil diameter during the observation of the critical events significantly predicted rating accuracy. As expected, the greater the pupil diameter, the less the raters deviated from the master ratings (ß = -0.15, $SE = 0.03$, $t = -4.83$, $p < .001$). For the aspect *challenging tasks*, the distribution of visual attention toward the teacher versus students was significant for rating accuracy. As expected, observers deviated less from the master ratings when they focused more on the teacher during critical events (ß = -0.05, $SE = 0.03$, $t = -1.71$, $p = .04$).

Additionally, we found some rating accuracy differences that depended on the video environment and subject-specific background. For the aspect *focus on key concepts*, observers with mathematics or physics backgrounds, who supposedly had more knowledge about the teaching content and subject-didactics, deviated less from the master ratings than observers with other subject backgrounds (ß = -0.37, $SE = 0.15$, $t = -2.40$, $p = .02$). For the aspect *challenging tasks*, observers deviated more from the master ratings in the immersive video environment (ß = 0.09, $SE = 0.04$, $t = 2.31$, $p = .02$). For the aspect *classroom disruptions*, observers deviated less from the master ratings in the immersive video environment (ß = -0.11, $SE = 0.04$, $t = -2.91$, $p < .01$).

**Table 4**

*Predicting Rating Accuracies From Event Focus, Eye-Tracking Measures, Video Environment, and Subject-Specific Background*

| | DMR: Focus on key concepts | | | | DMR: Challenging tasks | | | | DMR: Classroom disruptions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* |
| Intercept | 0.17 | 0.47 | 0.35 | .74 | 0.08 | 0.28 | 0.28 | .79 | -0.12 | 0.24 | -0.50 | .64 |
| Event focus[a] | 0.01 | 0.03 | 0.39 | .70 | -0.02 | 0.06 | -0.38 | .70 | 0.04 | 0.06 | 0.79 | .43 |
| Ratio of dwell time on teacher- and student-centered AOIs | 0.02 | 0.02 | 1.40 | .08* | **-0.05** | **0.03** | **-1.71** | **.04*** | -0.01 | 0.03 | 0.54 | .30* |
| Transitions between AOIs | -0.01 | 0.01 | -0.72 | .47 | 0.02 | 0.02 | 0.92 | .36 | 0.03 | 0.02 | 1.43 | .15 |
| Pupil diameter | **-0.15** | **0.03** | **-4.83** | **< .001*** | 0.06 | 0.05 | 1.08 | .14* | 0.02 | 0.04 | 0.51 | .31* |
| Video environment[b] | 0.01 | 0.02 | 0.41 | .68 | **0.09** | **0.04** | **2.31** | **.02** | **-0.11** | **0.04** | **-2.91** | **< .01** |
| Subject-specific background[c] | **-0.37** | **0.15** | **-2.40** | **.02** | 0.16 | 0.27 | 0.61 | .54 | 0.19 | 0.21 | 0.88 | .38 |

*Note.* DMR = deviation from master rating.

[a] 0 = teacher, 1 = students; [b] 0 = traditional, 1 = immersive; [c] 0 = no background in mathematics/physics, 1 = background in mathematics/physics.

*one-tailed testing based on a directional hypothesis

**Differences Between Video Environments**

With RQ3, we investigated whether the associations between event focus and gaze behavior (RQ1) and the associations between gaze behavior and rating accuracies (RQ2) differed between traditional and immersive video environments. Tables 5 and 6 present the interaction effects in the regression models for RQ1 and RQ2.

In predicting gaze behavior, we found one significant interaction between event focus and video environment, indicating that the association between event focus and pupil diameter was weaker in the immersive video environment (ß = -0.11, $SE = 0.04$, $t = -2.70$, $p < .01$; see Table 5).

In predicting rating accuracies, we found significant interactions between gaze-related indicators and video environment (see Table 6). For the aspect *challenging tasks*, we found that the significant association between the dwell time on the AOIs and rating accuracy was stronger in the immersive video environment (ß = 0.07, $SE = 0.03$, $t = 2.60$, $p < .01$). Regarding the teaching-quality aspects *focus on key concepts* and *classroom disruptions*, we found that the association between pupil diameter and rating accuracy was stronger in the immersive video environment than in the traditional video environment (*focus on key concepts*: ß = 0.07, $SE = 0.03$, $t = 2.80$, $p < .01$; *classroom disruptions*: ß = 0.11, $SE = 0.04$, $t = 2.79$, $p < .01$). As an example, Figure 3 illustrates the significant interaction between pupil diameter and video environment predicting the deviation from the master rating for the aspect *focus on key concepts*. Additionally, we found that for raters whose background matched the teaching content, the immersive video environment was not beneficial for rating the aspect *focus on key concepts*, as their ratings deviated more from the master ratings (ß = 0.35, $SE = 0.08$, $t = 4.49$, $p < .001$). Figure 4 illustrates this interaction.

**Table 5**

*Interactions With Video Environment in Predicting Observers' Gaze Behavior*

| | Ratio of dwell time on teacher- and student-centered AOIs | | | | Transitions between AOIs | | | | Pupil diameter | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* |
| Video Environment[b] × Event Focus[a] | 0.10 | 0.07 | 1.55 | .12 | 0.11 | 0.07 | 1.52 | .13 | **-0.11** | **0.04** | **-2.70** | **< .01** |
| Video Environment[b] × Subject-Specific Background[c] | -0.17 | 0.16 | -1.07 | .29 | 0.01 | 0.17 | 0.03 | .97 | -0.04 | 0.11 | -0.38 | .71 |

[a] 0 = teacher, 1 = students; [b] 0 = traditional, 1 = immersive; [c] 0 = no background in mathematics/physics, 1 = background in mathematics/physics.

**Table 6**

*Interactions With Video Environment in Predicting Rating Accuracies*

| | DMR: Focus on key concepts | | | | DMR: Challenging tasks | | | | DMR: Classroom disruptions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* | *ß* | SE | *t* | *p* |
| Video Environment × Event Focus | -0.04 | 0.03 | -1.36 | .18 | 0.07 | 0.05 | 1.36 | .17 | 0.05 | 0.05 | 0.89 | .38 |
| Video Environment × Ratio of Dwell Time on AOIs Teacher and Students | 0.03 | 0.02 | 1.61 | .11 | **0.07** | **0.03** | **2.60** | **< .01** | 0.02 | 0.03 | 0.86 | .39 |
| Video Environment × Transitions Between AOIs | <0.01 | 0.01 | 0.24 | .81 | -0.03 | 0.02 | -1.24 | .22 | <-.001 | 0.02 | -0.02 | .98 |
| Video Environment × Pupil Diameter | **0.07** | **0.03** | **2.80** | **< .01** | -0.05 | 0.04 | -1.14 | .25 | **0.11** | **0.04** | **2.79** | **< .01** |
| Video Environment × Subject-Specific Background | **0.35** | **0.08** | **4.49** | **< .001** | -0.16 | 0.13 | -1.21 | .23 | -0.12 | 0.13 | -0.95 | .34 |

*Note.* DMR = deviation from master rating.

[a] 0 = teacher, 1 = students; [b] 0 = traditional, 1 = immersive; [c] 0 = no background in mathematics/physics, 1 = background in mathematics/physics

**Figure 3**

*Significant Interaction Between Video Environment and Pupil Diameter Predicting Rating Accuracy*

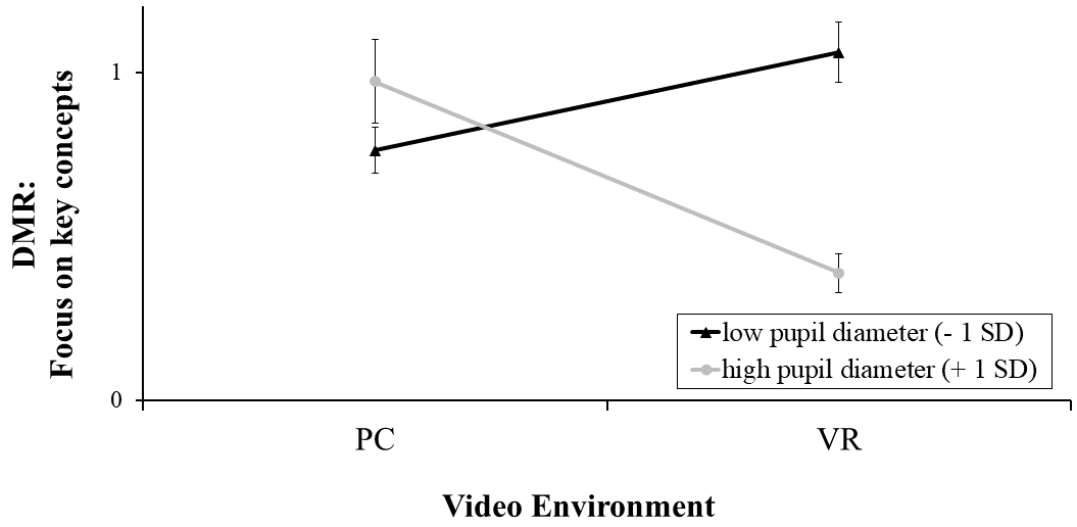
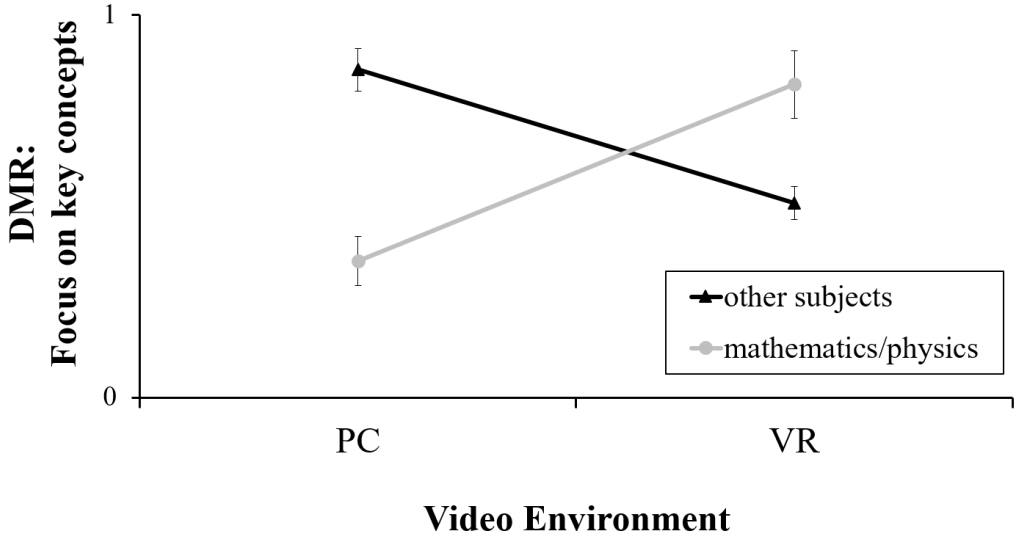



**Figure 4**

*Significant Interaction Between Video Environment and Observers' Subject-Specific Background Predicting Rating Accuracy*



## Discussion

The purpose of this study was to deepen the understanding of observer ratings of teaching quality in classroom videos by investigating cognitive processes during classroom observation using eye tracking. By linking gaze behavior to ratings of teaching quality, we gained valuable insights into the factors that contribute to explaining differences in teaching-quality ratings between observers and video environments.

### Findings

We found that different gaze behavior would result when critical events were conducted by the teacher versus the students (RQ1). Observers' dwell time on the teacher- and student-centered AOIs lined up with the event focus, supporting the idea that visual identification of the critical events may be a prerequisite for noticing these events (Just & Carpenter, 1976; Santagata et al., 2021). Additionally, we found evidence that, overall, observers focused more on the teacher in the immersive video environment and more on the students in the traditional video environment. This result contrasts with Kosko et al. (2021), who found a stronger focus on students' actions in immersive classroom video environments. There may be several reasons for the difference. First, Kosko et al. (2021) used written statements from observers rather than eye-tracking data. Additionally, variations in camera perspectives in the classroom could influence attention focus (Bozkir et al., 2021).

Investigating associations between gaze behavior and rating accuracy (RQ2), we found that some indicators of observers' gaze were related to rating accuracy. Ratings of *challenging tasks* were better aligned with master ratings when observers focused more on the teacher and less on the students. This finding supports our hypothesis, as the aspect *challenging tasks* is a more teacher-focused aspect of teaching quality (Fauth et al., 2022), thereby calling for a more teacher-focused observation to notice the critical events. Additionally, a larger pupil diameter was associated with more accurate ratings for the aspect *focus on key concepts*. We believe this result can be explained by the fact that observers who were both cognitively and emotionally more deeply involved in critical events (e.g., Appel et al., 2018; Goldinger & Papesh, 2012) may have more carefully considered the teaching, which could have improved their noticing and interpreting of critical events in order to provide accurate ratings.

In investigating whether associations in RQ1 and RQ2 depended on the complexity of information provided by the video environment (RQ3), we found evidence that associations between gaze behavior and rating accuracy were especially high in the immersive 360-degree video environment. A plausible explanation for this finding is that the immersive video

environment requires observers to actively explore and engage with the classroom scene, thereby demanding heightened attention and cognitive involvement (Haskins et al., 2020). Such increased engagement may facilitate more focused and comprehensive information processing, which could enhance the relationship between gaze behavior and rating accuracy. Furthermore, the immersive nature of 360-degree videos may minimize external distractions and amplify the sense of presence (Daltoè et al., 2024), contributing to a more accurate alignment between observers' gaze behavior and rating accuracy.

For observers with a subject-specific background in mathematics or physics, the immersive video environment seems to distract from the ability to observe the critical events and accurately assess teaching quality. We assume that observers with greater knowledge of the subject were better able to focus on the critical events and evaluate them more accurately for critical events in the traditional video environment. However, in an immersive 360-degree video environment, the additional complexity and richness of the surrounding information may divert their focus away from the critical events so that they focus on different areas in the 360-degree videos (Kosko et al., 2022).

**Limitations**

This study extensively advances current research on observer ratings of teaching quality. However, several limitations must be considered when interpreting the results. First, there are clear differences between the video environments, thus limiting their comparability. In the traditional video environment, classroom videos were presented in a split-screen format with teacher-focused and classroom-focused camera perspectives, where the whole classroom was presented to observers all the time. In the immersive video environment, however, participants observed the video from a student's seat in the second row, and it was necessary to move the head to observe the whole classroom. Furthermore, different eye-tracking systems were used between conditions. Another limitation is that AOIs were defined broadly and allowed us to differentiate between teacher and students but not between individual students or the exact part of the board. Furthermore, it is questionable whether seeing an event always means understanding it. This question might be particularly relevant for events that require a more subject-specific understanding of mathematics teaching (Schlesinger et al., 2018). Additionally, currently, our findings apply only to mathematics lessons and the specific events depicted in the videos. It would be beneficial to generalize these findings to other subjects and contexts.

**Implications and Future Research**

In this study, we used observers' gaze behavior during classroom observations in different video environments to understand what raters attend to in classroom videos and how this attention connects to rating quality. Our findings support the assumption that specific gaze-related indicators have the potential to predict the accuracy of teaching-quality ratings. This concept has important implications for the use of classroom observations in teacher education, school practice, and research.

First, rater training in classroom observations should address the visual focus of attention in the classroom video as well as cognitive attention to critical events during the observation as factors crucial for the success of observer ratings of teaching quality. For example, raters should be aware that assessing aspects of cognitive activation might be more accurate if they focused more closely on the teacher's behavior and if they actively attend to critical events relevant for this aspect of teaching quality. Second, using the right video environment with the right target group for the intended purpose of the classroom observation might be crucial. For example, a traditional video environment using a PC screen might be the right choice for a video-based reflection on aspects of cognitive activation for a class of mathematics students with high content knowledge. Third, this study highlights the significant potential of linking the noticing of critical classroom events to observer ratings of teaching quality. Because research on teacher noticing and research on teaching quality have traditionally been two well-established but separate fields, connecting them and building on each other's knowledge could enhance advancements in both areas.

Future research should compare traditional and 360-degree videos presented on the same technical device to make the conditions more comparable. Furthermore, studies should incorporate additional indicators of noticing and understanding critical events, such as open explanations for teaching-quality ratings (e.g., Praetorius et al., 2012). This approach could help show precisely which classroom events inform a rating. Moreover, the impact of raters' subject-specific background and their level of expertise should be investigated further in the context of teaching-quality ratings across different video environments (Dreher & Leuders, 2021). For instance, experienced teachers tend to focus more on students and distribute their gaze more evenly than inexperienced teachers (Keskin et al., 2024). These differences in visual attention may also correlate with variations in observer ratings of teaching quality.

In conclusion, this study underscores the potential of using eye-tracking technology to enhance the understanding of classroom observations. By capturing detailed gaze behavior,

eye-tracking offers valuable insights into the cognitive processes underlying how observers engage with and evaluate teaching practices. We encourage future research to build on these insights, utilizing eye tracking to further investigate the cognitive mechanisms involved in assessing teaching quality. Such studies will contribute to a deeper understanding of how observers interact with classroom videos, ultimately advancing knowledge of effective assessments of teaching quality.

## References

Appel, T., Scharinger, C., Gerjets, P., & Kasneci, E. (2018). Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Article 4, pp.1–8). Association for Computing Machinery. https://doi.org/10.1145/3204493.3204531

Atal, D., Admiraal, W., & Saab, N. (2023). 360° Video in teacher education: A systematic review of why and how it is used in teacher education. *Teaching and Teacher Education*, *135*, 104349. https://doi.org/10.1016/j.tate.2023.104349

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An Argument Approach to Observation Protocol Validity. *Educational Assessment*, *17*(2–3), 62–87. https://doi.org/10.1080/10627197.2012.715014

Bergin, C., Wind, S. A., Grajeda, S., & Tsai, C. L. (2017). Teacher evaluation: Are principals' classroom observations accurate at the conclusion of training?. *Studies in Educational Evaluation*, *55*, 19–26. https://doi.org/10.1016/j.stueduc.2017.05.002

Bozkir, E., Stark, P., Gao, H., Hasenbein, L., Hahn, J. U., Kasneci, E., & Göllner, R. (2021). Exploiting object-of-interest information to understand attention in VR classrooms. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (pp. 597–605). IEEE. http://dx.doi.org/10.1109/VR50410.2021.00085

Daltoè, T., Ruth-Herbein, E., Brucker, B., Jaekel, A. K., Trautwein, U., Fauth, B., Gerjets, P., & Göllner, R. (2024). Immersive insights: Unveiling the impact of 360-degree videos on pre-service teachers' classroom observation experiences and teaching-quality ratings. *Computers & Education*, *213*, 104976. https://doi.org/10.1016/j.compedu.2023.104976

Dreher, A., & Leuders, T. (2021). Fachspezifität von Unterrichtsqualität—aus der Perspektive der Mathematikdidaktik [Subject-specificity of instructional quality—From the perspective of mathematics education]. *Unterrichtswissenschaft*, *49*(2), 285–292. https://doi.org/10.1007/s42010-021-00116-9

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Herbein, E., & Maier, J. L. (2022). *Beobachtungsmanual zum Unterrichtsfeedbackbogen Tiefenstrukturen (2. aktualisierte Version) [Observation manual for the classroom feedback form deep structures (2. updated version)].* Institut für Bildungsanalysen Baden-Württemberg.

Gitomer, D. H. (2021). Methods for Observing Classroom Interaction. In R. Coe, M. Waring, L. Hedges, & L. Day Ashley (Eds.), *Research Methods and Methodologies in Education, 3rd Edn* (pp. 221–231). SAGE Publications.

Gold, B., & Windscheid, J. (2020). Observing 360-degree classroom videos–Effects of video type on presence, emotions, workload, classroom observations, and ratings of teaching quality. *Computers & Education*, *156*, 103960. https://doi.org/10.1016/j.compedu.2020.103960

Goldinger, S. D., & Papesh, M. H. (2012). Pupil Dilation Reflects the Creation and Retrieval of Memories. *Current Directions in Psychological Science*, *21*(2), 90–95. https://doi.org/10.1177/0963721412436811

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores. *American Journal of Education*, *119*(3), 445–470. https://doi.org/10.1086/669901

Grub, A. S., Biermann, A., Lewalter, D., & Brünken, R. (2022). Professional knowledge and task instruction specificity as influencing factors of prospective teachers' professional vision. *Teaching and Teacher Education*, *109*, 103517. https://doi.org/10.1016/j.tate.2021.103517

Haskins, A. J., Mentch, J., Botch, T. L., & Robertson, C. E. (2020). Active vision in immersive, 360° real-world environments. *Scientific Reports*, *10*(1), 14304. https://doi.org/10.1038/s41598-020-71125-4

Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *The Elementary School Journal*, *101*(1), 3–20.

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Janik, T., & Seidel, T. (2009). *The power of video studies in investigating teaching and learning in the classroom*. Waxmann.

Jentsch, A., Benecke, K., Blömeke, S., König, J., & Kaiser, G. (2024). Effects of observation mode on ratings of teaching quality in secondary mathematics classrooms. *ZDM*, *56*, 789–800. https://doi.org/10.1007/s11858-024-01557-z

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. https://doi.org/10.1016/0010-0285(76)90015-3

Keskin, Ö., Seidel, T., Stürmer, K., & Gegenfurtner, A. (2024). Eye-tracking research on teacher professional vision: A meta-analytic review. *Educational Research Review*, *42*, 100586. https://doi.org/10.1016/j.edurev.2023.100586

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Kosel, C., Holzberger, D., & Seidel, T. (2021). Identifying Expert and Novice Visual Scanpath Patterns and Their Relationship to Assessing Learning-Relevant Student Characteristics. *Frontiers in Education*, *5*, 612175. https://doi.org/10.3389/feduc.2020.612175

Kosko, K. W., Ferdig, R. E., & Zolfaghari, M. (2021). Pre-service Teachers' Professional Noticing When Viewing Standard and 360 Video. *Journal of Teacher Education*, *72*(3), 284–297. https://doi.org/10.1177/0022487120939544

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

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537. https://doi.org/10.1016/j.learninstruc.2008.11.001

McIntyre, N. A., Jarodzka, H., & Klassen, R. M. (2019). Capturing teacher priorities: Using real-world eye-tracking to investigate expert teacher priorities across two cultures. *Learning and Instruction*, *60*, 215–224. https://doi.org/10.1016/j.learninstruc.2017.12.003

Paulicke, P., Ehmke, T., Pietsch, M., & Schmidt, T. (2019). Wie beeinflusst die Kameraperspektive die Beurteilung der Unterrichtsqualität? [How does the camera perspective influence the assessment of teaching quality?]. *Zeitschrift für Bildungsforschung*, *9*(3), 411–435. https://doi.org/10.1007/s35834-019-00246-2

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom Assessment Scoring System [CLASS]: Manual, Pre-K. Brookes Publishing.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Santagata, R., König, J., Scheiner, T., Nguyen, H., Adleff, A.-K., Yang, X., & Kaiser, G. (2021). Mathematics teacher learning to notice: A systematic review of studies of video-based programs. *ZDM*, *53*(1), 119–134. https://doi.org/10.1007/s11858-020-01216-z

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM*, *50*(3), 475–490. https://doi.org/10.1007/s11858-018-0917-5

Stahnke, R., & Blömeke, S. (2021). Novice and expert teachers' noticing of classroom management in whole-group and partner work activities: Evidence from teachers' gaze and identification of events. *Learning and Instruction*, *74*, 101464. https://doi.org/10.1016/j.learninstruc.2021.101464

Stürmer, K., Seidel, T., Müller, K., Häusler, J., & Cortina, K. S. (2017). What is in the eye of pre-service teachers while instructing? An eye-tracking study about attention processes in different teaching situations. *Zeitschrift für Erziehungswissenschaft*, *20*(1), 75–92. https://doi.org/10.1007/s11618-017-0731-9

Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018). Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional Face Task. *Frontiers in Neurology*, *9*, 1029. https://doi.org/10.3389/fneur.2018.01029

## Appendix A

**Table A1**

*List of Critical Classroom Events in the Video Material*

| Video | Critical event | Event enacted by teacher/students | Event exemplifies this aspect of teaching quality |
|---|---|---|---|
| 1 | Student fidgets restlessly, waving their notebook. | Students | Classroom disruptions |
| | The student raises her hand exaggeratedly and becomes grumpy after not being called on. | Students | Classroom disruptions |
| | The students giggle loudly. | Students | Classroom disruptions |
| | The student raises her hand exaggeratedly again and becomes grumpy after not being called on. | Students | Classroom disruptions |
| | The teacher introduces a visualization of variables: "You can think of these as lengths of line segments. We call them *a* and *b* because we don't know how long they are, and they can change. Let's create a picture to illustrate this." | Teacher | Focus on key concepts |
| | Students throw paper balls while the teacher is drawing a line segment on the board. | Students | Classroom disruptions |
| | The teacher explains the visualization: "One side is *2*, and the other side is *3a* plus *2b*. That means 3 times line segment a, for example, this long (draws it), and the other line segment is 2 times line segment b, for example, this long (draws it)." | Teacher | Focus on key concepts |
| | Student is fidgeting restlessly in their chair. | Students | Classroom disruptions |
| | The teacher summarizes: "Just remember that with variables, we need to be aware that we don't know their values and that they can change." | Teacher | Focus on key concepts |

| | | | |
|---|---|---|---|
| 2 | The teacher introduces a visualization of variables: "You can think of these as lengths of line segments. We call them *a* and *b* because we don't know how long they are, and they can change. Let's create a picture to illustrate this." | Teacher | Focus on key concepts |
| | The teacher explains the visualization: "One side is *2*, and the other side is *3a* plus *2b*. That means 3 times line segment a, for example, this long (draws it), and the other line segment is 2 times line segment b, for example, this long (draws it)." | Teacher | Focus on key concepts |
| | The teacher summarizes: "Just remember that with variables, we need to be aware that we don't know their values and that they can change." | Teacher | Focus on key concepts |
| 3 | The teacher introduces a visualization of variables: "You can think of this like apples and pears. You can't just add them together either. Let's create a picture to illustrate this." | Teacher | Focus on key concepts |
| | The teacher draws apples and pears on the board. | Teacher | Focus on key concepts |
| | The teacher explains: "Variable *a* could stand for "apple," and variable *b* could stand for "pear" (in German: "**B**irne")." | Teacher | Focus on key concepts |
| | The teacher explains: "Because both terms describe the same fruit." | Teacher | Focus on key concepts |
| | The teacher summarizes: "Except that with variables, we have to be careful not to add apples and pears together." | Teacher | Focus on key concepts |
| 4 | The teacher introduces a third circle sector as an example. | Teacher | Challenging tasks |
| | A student responds to the task: "We need to know exactly how big the part of the full circle is." | Students | Challenging tasks |
| | The teacher hangs a sign with the degree measure on the board. | Teacher | Challenging tasks |
| | A student responds: "Yes, the full circle has 360 degrees." | Students | Challenging tasks |

| | | | |
|---|---|---|---|
| | A crucial clue for generalizing the formula to calculate the area of a circle sector comes from the students: “Aren’t these *185 out of 360*, like in the second example *3 out of 4*? … So you take the area of the full circle and then multiply it by *185/360*." | Students | Challenging tasks |
| | The teacher writes the equation on the board as dictated by the student. | Teacher | Challenging tasks |
| | A crucial clue for generalizing the formula to calculate the area of a circle sector comes from a student: "I think that with the circle sectors, which are either a three-quarter circle or a semicircle, it can also be done the same way using the *1/360th*. For a semicircle, this would simply be *180/360*." | Students | Challenging tasks |
| | The teacher writes another equation on the board as dictated by the student. | Teacher | Challenging tasks |
| 5 | The teacher writes the equation on the board as dictated by the student. | Teacher | Challenging tasks |
| | The teacher asks: "So, now we have gathered some examples. But if we have any arbitrary circle sector with any central angle alpha, how can we then calculate its area? Take another look at our examples on the board." | Teacher | Challenging tasks |
| | A student responds: "You can always calculate, for example, divided by 3 for a one-third circle, and thus calculate the area of any given sector of circle." | Students | Challenging tasks |
| | The teacher asks: "Can we derive a general formula from these ideas to calculate the area of any given circle sector? Take a look at the examples on the board again." | Teacher | Challenging tasks |
| | The teacher provides a crucial hint for generalizing the formula: "Yes, exactly. And for a semicircle, the central angle is 180 degrees. And *1/2* can also be written as *180/360* because the full circle has 360°, as we have already discussed." | Teacher | Challenging tasks |

## Appendix B

**Table B1**

*Correlations Between Metric Variables*

| | Ratio of Dwell Time on AOIs Teacher and Students | Transitions Between AOIs | Pupil Diameter | DMR: Focus on Key Concepts | DMR: Challenging Tasks | DMR: Classroom Disruptions |
|---|---|---|---|---|---|---|
| Ratio of Dwell Time on AOIs Teacher and Students | 1 | | | | | |
| Transitions Between AOIs | .01 | 1 | | | | |
| Pupil Diameter | .05 | **.09*** | 1 | | | |
| DMR: Focus on Key Concepts | .07 | **-.08**** | **-.24**** | 1 | | |
| DMR: Challenging Tasks | **-.07*** | .02 | **.09*** | **.09**** | 1 | |
| DMR: Classroom Disruptions | **-.09**** | .04 | **-.19**** | .03 | **-.10**** | 1 |

Note. DMR = Deviation from master rating.

* $p < .05$. ** $p < .01$.

# 6

# GENERAL DISCUSSION

# 6 General Discussion

The quality of classroom teaching is a crucial determinant of students' academic achievement and motivation (e.g., Blömeke et al., 2022; Burroughs et al., 2019). The construct of teaching quality has been widely conceptualized using three basic dimensions: cognitive activation, student support, and classroom management (Klieme et al., 2009; Pianta & Hamre, 2009). The most common approaches to assessing teaching quality include teachers' self-assessments, student ratings, and observer ratings based on classroom observations (Clausen, 2002; Fauth et al., 2020). Among these perspectives, observer ratings have been regarded as the gold standard (Helmke, 2009). Due to its potential to offer detailed insights into teaching practices, the observer perspective has been widely applied in teacher education and research, often utilizing classroom videos as representations of teaching practice (e.g., Janik & Seidel, 2009; van Es & Sherin, 2008). However, despite its promising opportunities, this perspective has notable conceptual challenges, as evidenced by research findings indicating limited psychometric quality (e.g., Kelly et al., 2020; Praetorius et al., 2012). To advance our understanding of observation-based assessment of teaching quality, research must systematically examine the conditions under which observers can accurately assess the quality of teaching from an external perspective. This dissertation aims to contribute to this endeavor by offering deeper insights into factors influencing the quality of observer ratings of teaching quality. My research is guided by a vision shared across efforts in this field: making teaching quality comprehensible to improve teaching and learning in our education system. Praetorius and Charalambous (2018, p. 535) summarize this overarching goal in their overview article on classroom observation frameworks for studying teaching quality:

> "Looking forward, we are convinced that we need to join efforts and intensify our attempts to learn from each other (…) and through that determine next steps on how we can better study, understand, and improve instructional quality."

To systematically investigate the assessment of teaching quality through classroom observation, I proposed a theoretical model of observers' assessment accuracy in classroom observations, adapted and extended from Funder's (1995) RAM, originally developed for personality judgement. This model provides a framework for examining the conditions under which observers can accurately assess teaching quality (see Figure 6, Chapter 1.4). It describes the process of assessing teaching quality through classroom observation and includes two factors influencing this process: the observers' disposition and the observation environment. To explore the proposed model and deepen our understanding of observer ratings of teaching

quality, Study 1 focused on how observer ratings develop in relation to the level of training, representing an aspect of observers' disposition. Studies 2 and 3 investigated the role of the observation environment by comparing traditional classroom video observations on a computer with immersive 360-degree video observations using VR glasses. Study 2 examined differences in observation experiences and absolute ratings of teaching quality, whereas Study 3 analyzed how observers' gaze behavior as process data of classroom observations is associated with observer ratings of teaching quality in different video environments.

In the following chapter, I summarize and discuss the findings of the three empirical studies in relation to the research questions of this dissertation and in the light of the proposed theoretical model of observers' assessment accuracy in classroom observations (Chapter 6.1). I then highlight the strengths and limitations of the dissertation (Chapter 6.2), explore implications for future research and practice (Chapter 6.3), and conclude with a general summary (Chapter 6.4).

## 6.1 Discussion of the Results

The present dissertation investigated three central research questions related to the assessment of teaching quality through classroom observations based on the proposed theoretical model underlying this dissertation. In the following chapter, I discuss the first research question and address the development of observer ratings over the course of a rater training based on the results of Study 1 (Chapter 6.1.1). Second, I will discuss the role of the video environment for classroom observations and assessments of teaching quality based on the results of Study 2 and Study 3 (Chapter 6.1.2). Finally, I will discuss the insights provided by using eye-tracking as process data of classroom observations based on the results of Study 3 (Chapter 6.1.3).

### 6.1.1 Development of Observer Ratings of Teaching Quality in a Rater Training

The first overarching research question of this dissertation examined how a rater training in classroom observation impacts observer ratings of teaching quality. I addressed this question in Study 1, which systematically investigated the development of observer ratings in the context of a rater training for in-service teachers participating in the pilot and validation study for a newly developed standardized classroom observation system for school practice in Baden-Württemberg, Germany (Unterrichtsfeedbackbogen Tiefenstrukturen (UFB); Fauth et al., 2021). All in all, the results of this study revealed that observers can assess teaching quality with an increasing level of agreement over the course of the rater training and that validity arguments support that ratings of trained teachers are valid measures of the intended teaching-quality aspects. These results support the assumption that the trained teachers developed an increasingly common understanding of the aspects of teaching quality assessed by the observation system (Bell et al., 2019; Joe et al., 2013; Klette, 2023; White & Ronfeldt, 2024). For this reason, this dissertation underscores the crucial role of rater training in reducing rater bias in classroom observations (e.g., Wang & Engelhard, 2019). It further demonstrates that rater bias is not a fixed disposition but can be diminished through targeted training. However, consistent with findings from previous research, the developments of rating agreement varied substantially between different aspects of teaching quality (Bell et al., 2014; Bergin et al., 2017). This made clear that some aspects of teaching quality barely need training to be rated with high psychometric quality, such as the aspects of the basic dimension of classroom management. On the other hand, there are aspects of teaching quality that require specific attention in training, such as the aspects of cognitive activation. The proposed model of

observers' assessment accuracy in classroom observation provides a promising approach to explain these findings.

Considering Research Question 1 in the light of the proposed model, the individual level of training is part of the observers' disposition to provide ratings of teaching quality. I assume that this disposition, in this case the level of training, may impact both the detection and utilization of relevant classroom events as evidence for their teaching-quality ratings. However, these steps are also impacted by the availability of relevant information in the first place. Therefore, reflecting on the results regarding the item-specific differences in rater agreement, it can be argued that there are large differences in the availability of information about different aspects of teaching quality (Fauth et al., 2020; Vazire, 2010). Whereas, for example, the information if the teaching is affected by classroom disruptions is usually directly available to the observer, the information on how challenging a specific task provided by the teacher is for the students, is less directly available and observers need to draw inferences to get to a reliable and valid assessment of this aspect of teaching quality (Lotz et al., 2013). Furthermore, the improved rating accuracy—indicating improved detection and utilization of information—can be explained using the proposed model. A key factor contributing to the improved detection and utilization of classroom events during rater training is that raters become increasingly familiar with the observation manual over the course of the training. The observation manual is a critical component of any standardized classroom observation system, providing detailed information about the rating criteria for each teaching-quality item included in the system (Bell et al., 2019; Praetorius & Charalambous, 2018). As raters gain familiarity with the behavioral indicators outlined for each item in the manual, they learn to identify classroom interactions that provide relevant evidence for assessing specific aspects of teaching quality. This knowledge likely enhances their ability to detect key classroom events and effectively utilize these observations to produce accurate numeric ratings of teaching quality. These findings, along with their interpretation through the theoretical model, offer practical implications for designing rater training in classroom observations (see Chapter 6.3.2).

### 6.1.2 Classroom Observation and Assessment of Teaching Quality in Different Video Environments

The second overarching research question of this dissertation examined how different video environments influence classroom observations and observer ratings of teaching quality. This question is addressed in Studies 2 and 3, both of which utilize data from a controlled lab study comparing traditional video and immersive 360-degree video environments for classroom

observation and teaching-quality assessment. Since this research question encompasses two components—the impact on the classroom observation experience and the observer ratings of teaching quality resulting from the classroom observation—the contribution of these studies regarding the second research question is twofold.

The first central contribution, addressed in Study 2, is a deeper understanding of how observers' self-reported observation experiences differ across various classroom video environments. The results highlight significant differences in how pre-service teachers perceive both the observed classroom teaching and the observation task itself. Key distinctions between video environments include the observers' level of involvement in the classroom situation and differences in motivation—both in terms of the effort invested in the observation task and the general enthusiasm for using classroom videos in teacher education. A frequently discussed concern in the context of immersive media is the potential for mental overload among users (Parong & Mayer, 2021; Roche et al., 2021; Sweller, 2011). However, Study 2 finds no significant differences in self-reported mental load between the video environments, suggesting that this potential risk may not be evident in the context of classroom observations using immersive 360-degree videos. Instead, the heightened involvement in the classroom situation and increased motivation for the observation task support the assumption that immersive 360-degree videos can provide pre-service teachers with a more realistic observation experience than traditional videos viewed on a computer screen. By offering pre-service teachers the sensation of being present in a classroom, immersive 360-degree videos may enable them to apply theoretical knowledge gained in pedagogical courses more effectively. This potential to bridge the gap between theory and practice in teacher education (Korthagen, 2007; McGarr et al., 2017) represents a promising strength of immersive 360-degree classroom videos.

The second central contribution is a deeper understanding of how teaching quality is assessed across different classroom video environments, as explored in Studies 2 and 3. Overall, the results indicate that no video environment is universally better suited for assessing teaching quality; instead, the suitability depends on the specific aspect of teaching quality being assessed. Importantly, the observed differences between video environments in teaching-quality ratings were consistently small in effect size. My dissertation studies focused on three aspects of teaching quality: *focus on key concepts*, *challenging tasks* (both assigned to the basic dimension of cognitive activation) and *classroom disruptions* (assigned to the basic dimension of classroom management). Results suggest that ratings for the aspect *focus on key concepts* and *classroom disruptions* align more closely with the intended teaching quality in immersive 360-degree video environments, whereas the aspect *challenging tasks* is rated more accurately in

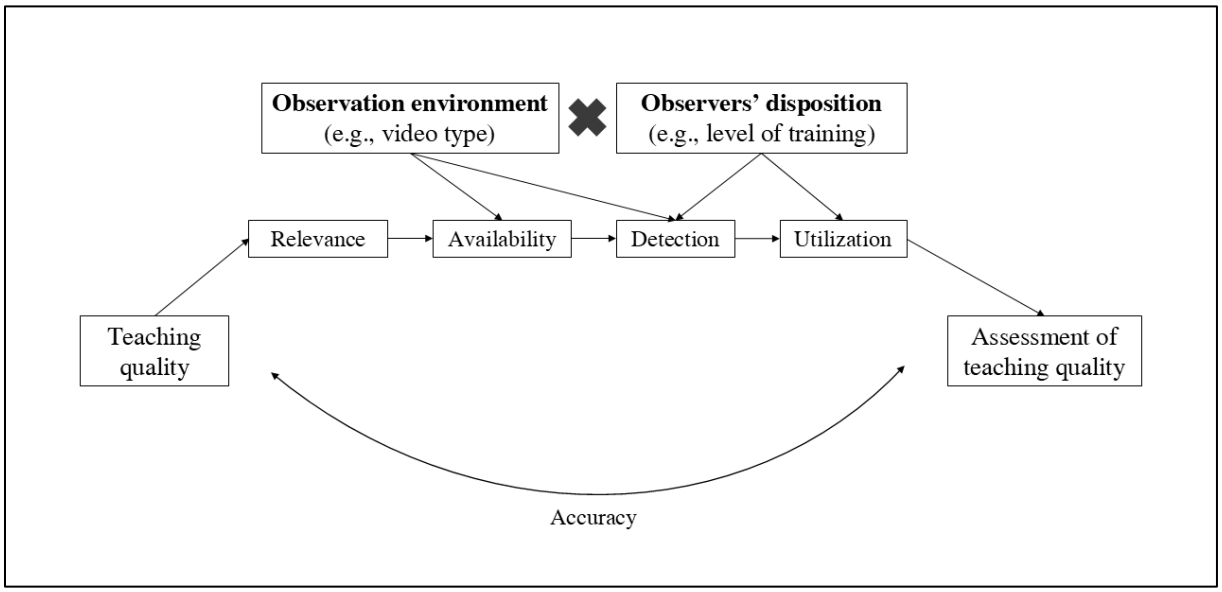
traditional video environments. This mixed pattern does not clearly separate more teacher-focused aspects of cognitive activation from the more student-focused aspect of classroom management, suggesting a complex interplay of video environment effects on the assessment of different teaching-quality aspects. When integrating the findings from both studies, a more nuanced pattern emerges. Evidence suggests that video environments may differentially affect the psychometric quality of observer ratings based on observers' subject-specific expertise and their visual and cognitive processes during classroom observation. For instance, pre-service teachers specializing in mathematics or physics—who are expected to be generally better equipped to accurately assess cognitive activation in mathematics lessons (Dreher & Leuders, 2021)—performed worse in the immersive video environment. This may be due to the increased complexity of information surrounding the observer in the immersive settings, which could distract observers from focusing on the instructional practices relevant to assessing the cognitive activation by the teacher. Conversely, observers with heightened attention and cognitive engagement in the classroom situation, indicated by pupil diameter, tended to perform particularly well in immersive environments (see Chapter 6.1.3).

Considering Research Question 2 in the light of the proposed model, comparing the video environment created by two different video types is part of the observation environment impacting classroom observations and resulting assessments of teaching quality. The research findings regarding Research Question 2 show that increasing the level of immersion in the observation environment by presenting a 360-degree classroom video does not directly result in an improved assessment of teaching quality. However, more importantly, the observation environment needs to provide the specific information that is relevant for an accurate assessment of teaching quality. When the immersive character of 360-degree videos supports the availability or detection of the relevant classroom events, this video environment should also support rating accuracy. However, when the immersive character of 360-degree videos harms the availability or detection of relevant classroom events, this video environment might also minimize rating accuracy. My results support the assumption that video environments have differential effects on the rating accuracy, depending on factors such as the observers' subject-specific background. Specifically, observers with a subject-specific background fitting to the teaching content were not able to use their knowledge to provide more accurate ratings in the immersive video environment. For this reason, I extend the proposed model of observers' assessment accuracy in classroom observations by incorporating an interaction between the observation environment and the observers' disposition (see Figure 8). This interaction highlights that the impact of both the observation environment and the observers' disposition

on the classroom observation assessment process is not uniform, but mutually influential. For example, the observers' disposition may affect the ability to integrate the complexity of information presented in different observation environments.

**Figure 8**

*Extension of the Proposed Model of Observers' Assessment Accuracy in Classroom Observations Based on the Dissertation Result*




## 6.1.3 Using Eye-Tracking as Process Data of Classroom Observations

The third overarching research question of this dissertation investigated the additional insights that eye-tracking data provide about classroom observations and observer ratings of teaching quality. This question is addressed in Study 3, which explores the relationship between observers' gaze behavior and their ratings of teaching quality. The use of eye-tracking in this study represents a significant methodological contribution to research on classroom observation and rating processes. Eye-tracking enables the examination of observers' visual focus of attention throughout classroom video observations. By analyzing whether observers attended to relevant AOIs during critical classroom events, Study 3 offers valuable insights into how their visual attention informs their ratings. Notably, a stronger focus on the teacher AOI was associated with more accurate ratings of the teacher-focused aspect *challenging tasks*. Beyond visual attention, eye-tracking also captures pupil diameter as an indicator of cognitive processing. In Study 3, larger pupil diameters—suggestive of heightened attention and cognitive engagement (e.g., Appel et al., 2018; Goldinger & Papesh, 2012; Wang et al., 2018)—were associated with more accurate ratings for the aspect *focus on key concepts*. However, the

interactions between gaze-related indicators and the video environment revealed that gaze behavior was particularly relevant for rating accuracy in immersive 360-degree video environments. In such environments, observers do not have a fixed angle of view where all relevant information is presented; instead, they must decide where to focus their attention. Therefore, the stronger relationship between gaze behavior and rating accuracy in immersive environments is highly plausible. Overall, the findings related to Research Question 3 demonstrate that observers' gaze behavior can be meaningfully connected to the accuracy of their teaching-quality ratings, particularly in immersive classroom video environments. Therefore, incorporating gaze-related indicators into classroom observation research has the potential to enhance our understanding of how teaching quality is assessed through classroom observations (see Chapter 6.3.1).

Considering Research Question 3 in the light of the proposed model, eye-tracking as a process-based measure from classroom observations allows us to examine whether the hypothesized processes underlying observers' assessment accuracy are also reflected in the observers' gaze. Specifically, eye-tracking provides insights into the detection of quality-relevant classroom events, a process referred to as "noticing" in research on teacher professional vision (König et al., 2022; Santagata et al., 2021; Seidel & Stürmer, 2014). Eye-tracking has been widely employed as a process-based method to investigate teachers' noticing skills and has proven to be highly informative (Grub et al., 2020; Keskin et al., 2024). For example, research on teacher professional vision has demonstrated differences in how expert and novice teachers allocate their visual attention in classroom settings. While expert teachers distribute their visual attention more evenly across the classroom, novices often focus narrowly on isolated and frequently irrelevant areas (Keskin et al., 2024). The theoretical assumptions underlying research on teacher professional vision align closely with those in the proposed model of observers' assessment accuracy in classroom observations. Given these parallels, I advocate for building on existing research on teacher professional vision to systematically investigate the conditions that enable observers to detect and effectively utilize quality-relevant classroom events to accurately assess teaching quality. Eye-tracking emerges as a particularly promising tool to support this endeavor. However, it needs to be further explored whether gaze-related indicators are only connected to rating accuracy in immersive 360-degree classroom videos, or if alternative eye-tracking measures or other process-based measures can also help explain rating accuracy in traditional classroom videos.

## 6.2 Strengths and Limitations

The three empirical studies I discussed above have distinct strengths and limitations that must be considered when interpreting their findings. First, this dissertation provides a strong conceptual contribution by introducing a theoretical model of observers' assessment accuracy in classroom observations. This proposed model outlines four steps of the process of assessing teaching quality through classroom observation. Moreover, it incorporates observers' dispositions and the observation environment as critical factors influencing the four-step assessment process. Whereas previous research has concluded that observer ratings of teaching quality exhibit limited psychometric quality and require critical consideration (Kelly et al., 2020; Praetorius et al., 2012; White & Klette, 2023), this new conceptual model now offers a framework for systematically investigating the conditions necessary to improve the psychometric quality of these ratings. By presenting this model, my dissertation advocates for more research into strategies for enhancing rating quality in classroom observations. The three studies included in this dissertation, each locatable in the proposed model, represent a strong contribution toward systematically exploring how to improve rating quality.

Another central strength of this dissertation lies in its methodological contributions to the investigation of observer ratings of teaching quality. First, the systematic investigation of a rater training in the context of the pilot study for a new standardized classroom observation system developed for teaching-quality feedback in school practice in Baden-Württemberg, Germany (Fauth et al., 2021) is a novel approach in using classroom observation in a practice context. New classroom observation systems, especially in the school-practice context, are often not developed and evaluated based on scientific standards (Tarkian et al., 2019). For this reason, providing a systematic approach to evaluating the reliability and validity of observer ratings for an instrument for school practice is a strong contribution of this dissertation. However, the data collection of Study 1, which presents the systematic investigation of rater training for teachers in school practice, has clear methodological limitations, as it is often the case with research in the field. The small sample size and the absence of a control group design do not allow for conclusions about the effectiveness of the rater training. Instead, the findings primarily describe the development of rating quality over the course of the rater training.

Another methodological contribution of this dissertation is the use of immersive 360-degree video environments as an innovative research tool in teaching-quality research. Whereas initial applications of 360-degree videos in teacher education focused on exploring teacher noticing (Kosko et al., 2021, 2022) or self-efficacy in classroom management (Atal et al., 2024),

employing 360-degree video as a research tool to examine the quality of observer ratings is a clear advancement in this field. Additionally, this dissertation introduces eye-tracking data as process data collected during classroom observations, marking another significant methodological contribution. Although eye-tracking has been widely applied in research on teacher professional vision (Grub et al., 2020; Keskin et al., 2024), prior research on teaching quality has primarily concentrated on observer ratings as the result of the observation process (e.g., Taut & Rakoczy, 2016; White, 2018). Some studies incorporated cognitive interviews to explore observers' thought processes and reasoning behind their ratings (Praetorius et al., 2012; Vinokic et al., 2024). However, using eye-tracking to generate process data during classroom observations is a novel approach in teaching-quality research. This method holds significant potential for providing deeper insights into the processes underlying observer ratings of teaching quality. By combining these methodological approaches—leveraging measures of rating accuracy from observer ratings research and eye-tracking methods established in teacher professional vision research—this dissertation bridges two previously separate, yet complementary fields. It highlights the need for further research at the intersection of classroom observation and familiar fields, such as teacher professional vision, to foster a more integrated understanding of the assessment of teaching quality through classroom observations.

The controlled lab study conducted to compare traditional and immersive classroom video environments in Studies 2 and 3 offers much more standardization than a study in the field, such as Study 1. However, besides the clear methodological contributions of Studies 2 and 3, this data collection comes with clear methodological limitations. First, the sample size of $N = 75$ pre-service teachers should be critically discussed. On one hand, the sample in this study is relatively large for this type of research, considering that many experimental studies on 360-degree classroom videos have used less than half of this sample size (Atal et al., 2024; Ferdig et al., 2023; Kosko et al., 2021, 2024; Walshe & Driver, 2019). On the other hand, a larger sample would have been needed to ensure sufficient statistical power to reliably detect small effects, such as differences in teaching-quality ratings between the video environments. Another methodological limitation of the lab study is that the two video conditions used are not fully comparable. Whereas the traditional videos were displayed on a computer screen, the 360-degree videos were shown using VR headsets. This research design was chosen to contrast an innovative approach to using classroom videos with the current practice. However, with this setup, the two video conditions differed not only in terms of video type but also in the display device, introducing an additional confounding factor. This type of research design, often referred to as media comparison studies, has been widely used over the past decades to

investigate the potential of innovative educational technologies, but it has increasingly been criticized due to these additional confounding factors (Buchner & Kerres, 2023; Lawson et al., 2024). Another limitation of the lab study concerns the measurement of the observation experience investigated in Study 2. This study included the cognitive, affective, and physiological involvement in the classroom situation, the motivation by the classroom video, as well as mental load and mental effort as different aspects of the observers' experiences during classroom observation. These aspects were all measured via participants' self-reported survey data. However, given a potential bias inherent in self-reported data (e.g., Moore & Picou 2018), integrating multimodal measures of cognitive, affective, physiological, and motivational processes could have yielded a more robust and comprehensive understanding of the classroom observation experience. (Chen et al., 2016; Dubovi, 2022; Ferdig et al., 2023; Noroozi et al., 2020; Sümer et al., 2023; Vanneste et al., 2021).

Another significant strength of this work is the interdisciplinary character of this dissertation project. In this dissertation, a set of staged mathematics classroom videos was developed and produced in collaboration with practitioners (teachers, who incorporated their practical experience) and researchers from mathematics subject-didactics as well as generic teaching-quality research. These videos depict teaching situations based on scientifically grounded, didactically relevant situations, presenting quality differences regarding specific aspects of teaching quality assessed by the observation system by Fauth et al. (2021). Five of these staged classroom videos depicting quality differences for three aspects of teaching quality were used as stimulus material for the data collection for Study 2 and Study 3. On the one hand, the presentation of staged videos, including carefully selected teaching practices that demonstrate specific aspects of teaching quality has clear advantages for a standardized and theoretically grounded investigation of teaching-quality assessment (Seidel et al., 2022). On the other hand, the results about teaching-quality assessment from Study 2 and Study 3 are limited to the three focused aspects of teaching quality and there was no experimental manipulation of the availability of relevant classroom events between video environments, which might have allowed to test more targeted hypothesis about potential benefits of immersive 360-degree videos for the assessment of teaching quality.

The different limitations of this dissertation result in several implications for future research. The respective implications are presented in Chapter 6.3.1.

## 6.3 Implications and Future Directions

Building on the discussion of my dissertation's research questions and the overall strengths and limitations of this dissertation, the following chapter will explore implications and future research directions. In Chapter 6.3.1, I will outline implications for future research, followed by implications for practice in Chapter 6.3.2.

### 6.3.1 Implications for Future Research

In the present dissertation, I proposed a theoretical model of observers' assessment accuracy in classroom observation and, based on aspects of this model, I addressed three different approaches to get deeper insights into conditions that impact the observation-based assessment of teaching quality. The dissertation's results and the proposed theoretical model have several implications for future research.

Study 1 of this dissertation systematically investigated the development of rating quality over the course of rater training in classroom observation and derived different validity arguments for the use of the trained teachers' ratings after the training. The derived validity arguments indicate that the observer ratings of trained teachers can be used as a valid assessment of the intended teaching-quality aspects in future studies. This implication is limited to the specific classroom observation system designed for teaching-quality development in school practice in Baden-Württemberg, Germany (Unterrichtsfeedbackbogen Tiefenstrukturen (UFB); Fauth et al., 2021) and the accompanying rater training provided by the Zentrum für Schulqualität und Lehrerbildung (ZSL), that was investigated in this dissertation. As this classroom observation system has been broadly rolled out in the education system in the German state of Baden-Württemberg (Ruth-Herbein et al., 2022), the UFB, as well as the accompanying rater training will be widely applied for teaching-quality development in school practice. To gain robust results on the effectiveness of the rater training, future research should use advanced research designs, such as randomized control-group designs where different aspects of the rater training, such as the duration, training elements or training material, are experimentally varied and effects on the rating accuracy are investigated (Joe et al., 2013). Furthermore, the literature on the effectiveness of FOR training implies that feedback on the rating performance increases the effectiveness of the training (e.g., Elder et al., 2005). For this reason, investigating the impact of feedback on the individual rating performance throughout the rater training would be a promising approach to potentially improve the trained teachers' rating quality further. Besides investigating the effectiveness of the rater training itself, an

important question is whether the ratings of the trained teachers fulfill their ultimate goal to provide useful feedback on teaching quality for teaching-quality development in school practice (Ruth-Herbein et al., 2022). For this reason, the transfer of the acquired rating skills to practice (Baldwin & Ford, 1988) and their usefulness for developing teaching quality in school practice (Bell & Gitomer, 2023) would be another highly relevant field for future research.

Study 2 of this dissertation compared classroom observation experiences and observer ratings of teaching quality between a traditional and an immersive 360-degree classroom video environment. As highlighted in the previous chapter, future research should prioritize capturing classroom observation experiences using multimodal data (Chen et al., 2016; Dubovi, 2022; Ferdig et al., 2023; Noroozi et al., 2020; Sümer et al., 2023; Vanneste et al., 2021). This includes investigating whether objective data sources confirm the patterns observed in studies relying on subjective survey data, particularly regarding factors such as involvement in the classroom situation or mental load. Regarding the assessment of teaching quality, this dissertation demonstrated that immersive video environments are neither universally superior nor inferior. Instead, I found differential effects, depending on the assessed aspect of teaching quality and rater characteristics. Future research should build on these findings by formulating targeted hypotheses about which aspects of teaching quality are expected to be observed and rated more accurately in immersive environments and which aspects might be hindered by such settings. To test these hypotheses, researchers should select videotaped classroom situations designed to evaluate these theoretical predictions. Additionally, future research on immersive 360-degree classroom videos should extend beyond the assessment of teaching quality as target variable of classroom observation. It is important to explore for which other target variables immersive environments and the heightened sense of involvement they afford may be beneficial. For instance, do observers perceive classroom information with greater specificity? Can pre-service teachers memorize observed teaching practices more effectively and apply them in their own teaching? These and other questions addressing various target variables of immersive classroom video observation require further investigation. Moreover, research findings should be integrated into teacher training programs to make use of the potential benefits of immersive classroom video environments effectively.

Study 3 of this dissertation investigated associations between observers' gaze behavior and their ratings of teaching quality in a traditional and an immersive 360-degree classroom video environment. Future research should expand this research approach by using a variety of gaze-related indicators of observers' cognitive processes of classroom observation, such as scan paths, that reflect sequences of fixations and saccades within a given time period (e.g., Kosel et

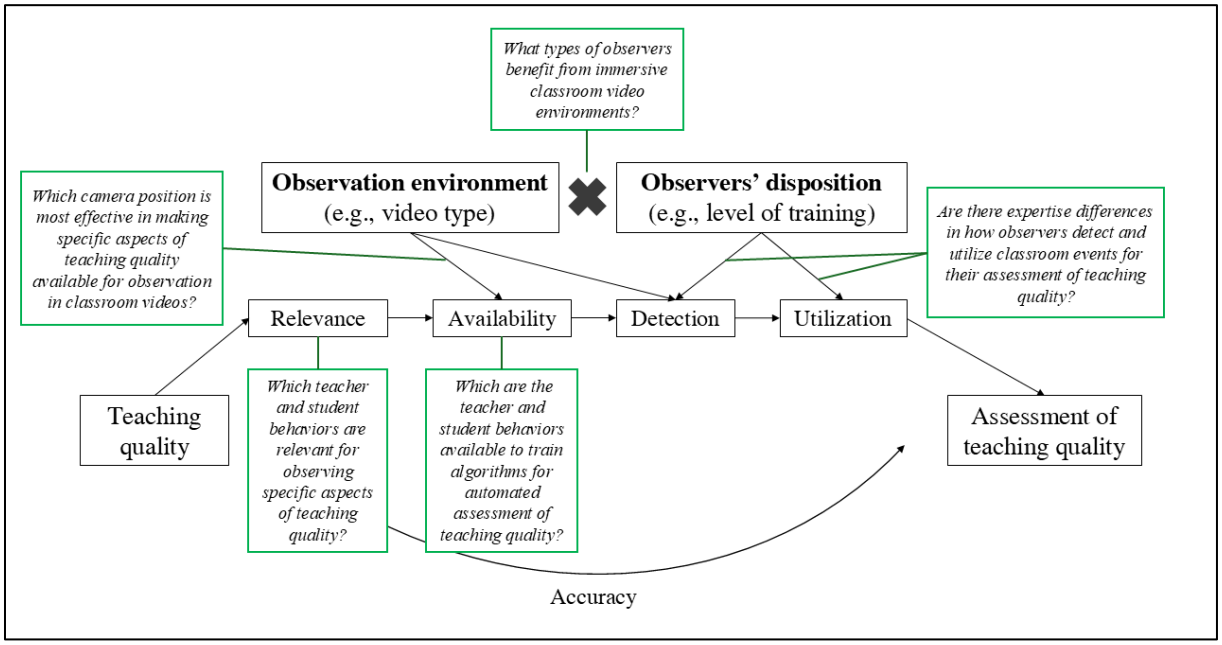
al., 2021), or the Gini coefficient, that is used as a measure of uneven gaze distribution between different AOIs (e.g., Cortina et al., 2015). Building on findings from research on teacher professional vision (Keskin et al., 2024), future research on the observation-based assessment of teaching quality should also consider the role of observers' expertise regarding the detection and utilization of quality-relevant classroom events in both traditional and immersive 360-degree classroom videos. Additionally, using further data sources on what observers noticed and utilized for their assessments of teaching quality based on classroom videos, for example through think-aloud protocols or cognitive interviews (e.g., Praetorius et al., 2012; Vinokic et al., 2024), would be a beneficial addition to eye-tracking data from classroom observations in future studies. A combination of different methods to shed light on the observers' cognitions would create a more comprehensive picture of the underlying cognitive processes involved in assessing teaching quality through classroom observation.

In addition to the implications for future research derived from the dissertation's results, I will take a closer look at specific implications stemming from the proposed model of observers' assessment accuracy in classroom observations. The adaptation and extension of Funder's (1995) RAM from the context of personality judgment to the context of teaching-quality assessment calls for a systematic consideration of the individual components of the model. For instance, the model highlights the importance of ensuring that behavioral cues relevant to teaching quality are available to observers assessing teaching quality. One potential way to enhance the availability of such cues in an immersive classroom video environment is by allowing observers to switch between different camera positions. This capability could heighten observers' perception of initiating and controlling their actions within the classroom environment, a phenomenon referred to as the sense of agency (Braun et al., 2018). An increased sense of agency in immersive environments, in turn, has been shown to elevate germane cognitive load, which refers to the cognitive resources dedicated to processing content (Lehikko et al., 2024; Li et al., 2023). Consequently, incorporating agency into immersive classroom video environments may influence the accuracy of teaching-quality assessments by improving both the availability of relevant information and the germane cognitive load associated with the observation task. In addition to this example regarding availability, future research should further investigate what can be learned from each step in the proposed model—relevance, availability, detection, and utilization—as well as the additional elements of observers' dispositions, the observation environment, and their interaction, to enhance the accuracy of teaching quality assessments. As an initial step to facilitate this, Figure 9 presents

exemplary research questions within the proposed model of observers' assessment accuracy in classroom observations (highlighted in the green-bordered boxes).

**Figure 9**

*Exemplary Research Questions for Future Research Within the Proposed Model of Observers' Assessment Accuracy in Classroom Observations*




## 6.3.2 Implications for Practice

Besides the implications for future research, the findings of this dissertation have significant practical implications for the application of classroom observation and observer ratings of teaching quality. The first area for practical implications concerns using the UFB (Fauth et al., 2021) as a standardized classroom observation system for teaching quality development in school practice, as well as the accompanying rater training for teachers provided by the ZSL. The validity arguments derived from the first study in this dissertation confirm that ratings provided by trained teachers are valid and offer valuable feedback on the aspects of teaching quality assessed by the UFB. Using these ratings for mutual feedback among teachers, for self-reflection on teaching quality, or as a tool for discussion and learning about teaching quality in teacher education and professional development programs can foster a common language around teaching quality (Klette, 2023) based on the TBD, and help improve teaching quality in school practice. Furthermore, the insights into the rating development throughout the rater training can be used to refine the training design. It became evident that some aspects of teaching quality in the observation system seem to require less training to achieve high

agreement among raters and with master ratings. These aspects include the rating items of classroom management and the more emotional-affective aspects of student support, such as classroom climate[2]. Rater training should address these aspects, but no in-depth training seems to be necessary for teachers to observe them and apply the corresponding scoring rules to generate reliable and valid ratings. On the other hand, there are aspects of teaching quality in the observation system which proved to be more challenging for teachers to observe and assess consistently. These include aspects of cognitive activation and more instructional elements of student support, such as the aspects *quality of feedback* or *scaffolding*. Rater training should offer a deep engagement with these more challenging aspects of teaching quality and strongly focus on the observable indicators that help assess these aspects accurately. Additionally, the importance of the visual focus of attention in the classroom (videos) is another implication derived from my dissertation's findings. As the results show, the accuracy of ratings can depend on where the observer directs their attention in a classroom video, for example, focusing on the teacher and the board when assessing the difficulty of a presented task. Incorporating this knowledge into rater training by providing guidance on where to focus attention when rating specific aspects of teaching quality could help observers direct their attention to the relevant areas in the classroom.

The second area of practical implications concerns using immersive 360-degree videos in teacher education and professional development programs (e.g., Atal et al., 2023). This dissertation provides strong evidence of the benefits of this video format, particularly in terms of increased motivation and its ability to immerse observers in a realistic classroom environment. Teacher educators could leverage these advantages by offering video-based training programs focusing on developing professional competencies and enhancing teaching quality through immersive 360-degree videos. Such an approach could spark the interest of (pre-service) teachers in participating in these courses. By creating supportive instructional settings (e.g., through shared discussions of the observed teaching), teacher educators could harness the increased motivation generated by immersive classroom observations to deepen the observers' understanding of teaching quality. In these contexts, the realistic classroom setting provided by immersive 360-degree videos may serve as an effective environment to bridge the theory-practice gap in teacher education (Korthagen, 2007; McGarr et al., 2017).

[2] These results regarding specific aspects of teaching quality refer to the items of the Unterrichtsfeedbackbogen Tiefenstrukturen (Fauth et al., 2021), which was piloted and validated in Study 1 of this dissertation. For other standardized classroom observation systems that focus on different aspects of classroom management and student support, the findings may vary.

## 6.4 Conclusion

In the introduction to this dissertation, we heard that Mrs. Cortes' Spanish class left a lasting positive impression. Beyond such personal impressions of good and bad teaching, making informed statements about teaching quality—particularly for the application in teacher education and research—requires accurate methods to capture this construct. One such method is classroom observation by external observers. While the observer perspective on teaching quality has been regarded as the gold standard for assessing teaching quality (Helmke, 2009), it also presents clear challenges concerning the accuracy of such assessments. This dissertation has contributed to a deeper understanding of the assessment of teaching quality through classroom observation. It proposed a theoretical model for observers' assessment accuracy in classroom observations and introduced three innovative approaches that advance teaching-quality research and the practical application of video-based classroom observations in teacher education and professional development. The findings emphasize the importance of monitoring rating quality during teacher training in classroom observation to understand the development of rating performance better. Furthermore, this work highlights the potential of immersive 360-degree classroom videos as an innovative tool for teaching-quality research and teacher training in realistic classroom settings. While these immersive videos offer unique advantages, they are not inherently superior for assessing teaching quality; instead, the observation environment must align with the specific purpose of the assessment. Finally, this dissertation demonstrated the significant value of eye-tracking technology in moving beyond observer ratings to investigate the underlying perceptual processes that shape assessments of teaching quality.

Overall, the present dissertation advanced the understanding of teaching-quality assessments through classroom observation in theoretical, methodological, and empirical dimensions. It not only serves as a comprehensive exploration of the observer perspective on teaching quality, but also establishes a strong foundation for future research endeavors that integrate theoretical and methodological approaches from related research fields. Building on these efforts will further enhance our ability to assess, understand, and ultimately improve teaching quality, ensuring that learners receive the most effective teaching possible.

# 7

# REFERENCES

# 7 References

Adhanom, I. B., MacNeilage, P., & Folmer, E. (2023). Eye Tracking in Virtual Reality: A Broad Review of Applications and Challenges. *Virtual Reality*, *27*(2), 1481–1505. https://doi.org/10.1007/s10055-022-00738-z

Ainley, J., & Carstens, R. (2018). *Teaching and Learning International Survey (TALIS) 2018 Conceptual Framework.* OECD Education Working Papers Series, No. 187, OECD Publishing, Paris. http://dx.doi.org/10.1787/799337c2-en

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, *39*(17), 2947–2953. https://doi.org/10.1016/S0042-6989(99)00019-X

Appel, T., Scharinger, C., Gerjets, P., & Kasneci, E. (2018). Cross-subject workload classification using pupil-related measures. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 1–8. https://doi.org/10.1145/3204493.3204531

Ardisara, A., & Fung, F. M. (2018). Integrating 360° Videos in an Undergraduate Chemistry Laboratory Course. *Journal of Chemical Education*, *95*(10), 1881–1884. https://doi.org/10.1021/acs.jchemed.8b00143

Atal, D., Admiraal, W., & Saab, N. (2023). 360° Video in teacher education: A systematic review of why and how it is used in teacher education. *Teaching and Teacher Education*, *135*, 104349. https://doi.org/10.1016/j.tate.2023.104349

Atal, D., Admiraal, W., & Saab, N. (2024). Effects of 360 ° video virtual reality-supported reflection on student teachers' classroom management self-efficacy and their stress levels. *Teaching and Teacher Education*, *144*, 104573. https://doi.org/10.1016/j.tate.2024.104573

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*(1), 63–105. https://doi.org/10.1111/j.1744-6570.1988.tb00632.x

Baumert, J., & Köller, O. (2000). Unterrichtsgestaltung, verständnisvolles Lernen und multiple Zielerreichung im Mathematik- und Physikunterricht der gymnasialen Oberstufe [Lesson design, learning with understanding and multiple goal achievement in mathematics and physics lessons in the upper secondary school]. In J. Baumert, W. Bos

& R. Lehmann (Eds.), *TIMSS III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie - Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Bd. 2. Mathematische und physikalische Kompetenzen in der Oberstufe* (pp. 271–315). Leske + Budrich.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal*, *47*(1), 133–180. https://doi.org/10.3102/0002831209345157

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O., & Neubrand, J. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde [TIMSS - International Comparison of Mathematics and Science Teaching: Descriptive Findings]*. Leske + Budrich.

Begrich, L., Kuger, S., Klieme, E., & Kunter, M. (2021). At a first glance – How reliable and valid is the thin slices technique to assess instructional quality? *Learning and Instruction*, *74*, 101466. https://doi.org/10.1016/j.learninstruc.2021.101466

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Bell, C. A., & Gitomer, D. H. (2023). Building the field's knowledge of teaching and learning: Centering the socio-cultural contexts of observation systems to ensure valid score interpretation. *Studies in Educational Evaluation*, *78*, 101278. https://doi.org/10.1016/j.stueduc.2023.101278

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An Argument Approach to Observation Protocol Validity. *Educational Assessment*, *17*(2–3), 62–87. https://doi.org/10.1080/10627197.2012.715014

Bell, C. A., Qi, Y., Croft, A. J., & Leusner, D. (2014). Improving Observational Score Quality: Challenges in Observer Thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 50–97). Jossey-Bass.

Bergin, C., Wind, S. A., Grajeda, S., & Tsai, C.-L. (2017). Teacher evaluation: Are principals' classroom observations accurate at the conclusion of training? *Studies in Educational Evaluation*, *55*, 19–26. https://doi.org/10.1016/j.stueduc.2017.05.002

Berliner, D. C. (2005). The Near Impossibility of Testing for Teacher Quality. *Journal of Teacher Education*, *56*(3), 205–213. https://doi.org/10.1177/0022487105275904

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in Rater Training. *Academy of Management Review*, *6*(2), 205–212. https://doi.org/10.5465/amr.1981.4287782

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, *79*, 101600. https://doi.org/10.1016/j.learninstruc.2022.101600

Blömeke, S., & Olsen, R. V. (2019). Consistency of results regarding teacher effects across subjects, school levels, outcomes and countries. *Teaching and Teacher Education*, *77*, 170–182. https://doi.org/10.1016/j.tate.2018.09.018

Bloom, B. S. (1976). *Human characteristics and school learning.* McGraw Hill.

Böheim, R., Urdan, T., Knogler, M., & Seidel, T. (2020). Student hand-raising as an indicator of behavioral engagement and its role in classroom learning. *Contemporary Educational Psychology*, *62*, 101894. https://doi.org/10.1016/j.cedpsych.2020.101894

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*(4), 410–421. https://doi.org/10.1037/0021-9010.64.4.410

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). The Impact of Assessment and Accountability on Teacher Recruitment and Retention: Are There Unintended Consequences? *Public Finance Review*, *36*(1), 88–111. https://doi.org/10.1177/1091142106293446

Braun, N., Debener, S., Spychala, N., Bongartz, E., Soros, P., Müller, H. H. O., & Philipsen, A. (2018). The senses of agency and ownership: A review. *Frontiers in Psychology*, *9*, 535. https://doi.org/10.3389/fpsyg.2018.00535

Brophy, J. (2000). *Teaching*. Educational Practices Series-1, International Academy of Education.

Brophy, J. (2004). *Using video in teacher education.* Elsevier.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). Macmillan.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217.

Brunvand, S. (2010). Best practices for producing video content for teacher education. *Contemporary Issues in Technology and Teacher Education, 10*(2), 247–256.

Buchner, J., & Kerres, M. (2023). Media comparison studies dominate comparative research on augmented reality in education. *Computers & Education*, *195*, 104711. https://doi.org/10.1016/j.compedu.2022.104711

Bühler, B., Hou, R., Bozkir, E., Goldberg, P., Gerjets, P., Trautwein, U., & Kasneci, E. (2023). Automated Hand-Raising Detection in Classroom Videos: A View-Invariant and Occlusion-Robust Machine Learning Approach. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 102–113). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_9

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., & Schmidt, W. (2019). A Review of the Literature on Teacher Effectiveness and Student Outcomes. In N. Burroughs, J. Gardner, Y. Lee, S. Guo, I. Touitou, K. Jansen, & W. Schmidt (Eds.), *Teaching for Excellence and Equity* (pp. 7–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-16151-4_2

Calvert, J., & Abadia, R. (2020). Impact of immersing university and high school students in educational linear narratives using virtual reality technology. *Computers & Education*, *159*, 104005. https://doi.org/10.1016/j.compedu.2020.104005

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(6), 723–733.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of Observation Mode on Measures of Secondary Mathematics Teaching. *Educational and Psychological Measurement*, *73*(5), 757–783. https://doi.org/10.1177/0013164413486987

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly*, *27*(3), 529–542. https://doi.org/10.1016/j.ecresq.2011.12.006

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust Multimodal Cognitive Load Measurement*. Springer International Publishing. https://doi.org/10.1007/978-3-319-31700-7

Ciesielska, M., Boström, K. W., & Öhlander, M. (2018). Observation Methods. In M. Ciesielska & D. Jemielniak (Eds.), *Qualitative Methodologies in Organization Studies* (pp. 33–52). Springer International Publishing. https://doi.org/10.1007/978-3-319-65442-3_2

Clare, L., Valdes, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (CSE Technical Report No. 545). National Center for Research on Evaluation, CRESST/CSE, Graduate School of Education & Information Studies, University of California, Los Angeles. https://eric.ed.gov/?id=ed457169

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstukt- und Kriteriumsvalidität [Teaching quality: A question of perspective? Empirical analyses of agreement, construct and criterion validity]*. Waxmann.

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (2020). Between authenticity and cognitive demand: Finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teaching and Teacher Education*, *95*, 103146. https://doi.org/10.1016/j.tate.2020.103146

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Government Printing Office.

Cortina, K. S., Miller, K. F., McKenzie, R., & Epstein, A. (2015). Where Low and High Inference Data Converge: Validation of CLASS Assessment of Mathematics Instruction Using Mobile Eye Tracking with Expert and Novice Teachers. *International Journal of Science and Mathematics Education*, *13*(2), 389–403. https://doi.org/10.1007/s10763-014-9610-5

Cortina, K. S., Müller, K., Häusler, J., Stürmer, K., Seidel, T., & Miller, K. F. (2018). Feedback mit eigenen Augen: Mobiles Eyetracking in der Lehrerinnen- und Lehrerbildung [Feedback with your own eyes: Mobile eye tracking in teacher education]. *Beiträge zur Lehrerinnen- und Lehrerbildung, 36*(2), 208–222. https://doi.org/10.25656/01:17097

Curby, T. W., Johnson, P., Mashburn, A. J., & Carlis, L. (2016). Live Versus Video Observations: Comparing the Reliability and Validity of Two Methods of Assessing Classroom Quality. *Journal of Psychoeducational Assessment*, *34*(8), 765–781. https://doi.org/10.1177/0734282915627115

Danielson, C. (2007). *Enhancing professional practice*. Association for Supervision and Curriculum Development.

Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Dreher, A., & Leuders, T. (2021). Fachspezifität von Unterrichtsqualität—aus der Perspektive der Mathematikdidaktik [Subject-specificity of instructional quality—From the perspective of mathematics education]. *Unterrichtswissenschaft*, *49*(2), 285–292. https://doi.org/10.1007/s42010-021-00116-9

Dubovi, I. (2022). Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology. *Computers & Education*, *183*, 104495. https://doi.org/10.1016/j.compedu.2022.104495

Dumont, H. (2019). Neuer Schlauch für alten Wein? Eine konzeptuelle Betrachtung von individueller Förderung im Unterricht [Adaptive teaching: Conceptual reflections]. *Zeitschrift für Erziehungswissenschaft*, *22*(2), 249–277. https://doi.org/10.1007/s11618-018-0840-0

Elder, C., Knoch, U., Barkhuizen, G., & Von Randow, J. (2005). Individual Feedback to Enhance Rater Training: Does It Work? *Language Assessment Quarterly*, *2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Erickson, F. (2007). Ways of seeing video: Toward a phenomenology of viewing minimally edited footage. In R. Goldman, P. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 145–155). Erlbaum.

Evens, M., Empsen, M., & Hustinx, W. (2023). A literature review on 360-degree video as an educational tool: Towards design guidelines. *Journal of Computers in Education*, *10*(2), 325–375. https://doi.org/10.1007/s40692-022-00233-z

Fauth, B., Decristan, J., Rieser, S., & Klieme, E. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg [Teaching quality in primary school from the perspectives of students, teachers, and external observers: Relationships between perspectives and prediction of student achievement]. *Zeitschrift für Pädagogische Psychologie*, *28*(3), 127–137. https://doi.org/10.1024/1010-0652/a000129

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who Sees What?: Conceptual Considerations on the Measurement of Teaching Quality from Different Perspectives. *Zeitschrift für Pädagogik Beiheft*, *1*, 138–155. https://doi.org/10.3262/ZPB2001138

Fauth, B., Herbein, E., & Maier, J. L. (2021). Beobachtungsmanual zum Unterrichtsfeedbackbogen Tiefenstrukturen [Observation manual for the classroom feedback form deep structures]. Institut für Bildungsanalysen Baden-Württemberg.

Ferdig, R. E., & Kosko, K. W. (2020). Implementing 360 Video to Increase Immersion, Perceptual Capacity, and Teacher Noticing. *TechTrends*, *64*(6), 849–859. https://doi.org/10.1007/s11528-020-00522-3

Ferdig, R., Kosko, K. W., & Gandolfi, E. (2023). Using Fitbits and heart rate variance (HRVa) to understand pre-service teacher experiences in extended reality. In E. Langran, P. Christensen, & J. Sanson (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 1173–1179). New Orleans, LA, United States: Association for the Advancement of Computing in Education (AACE). Retrieved December 10, 2024 from https://www.learntechlib.org/primary/p/221982/

Foster, J. K., Korban, M., Youngs, P., Watson, G. S., & Acton, S. T. (2024). Automatic classification of activities in classroom videos. *Computers and Education: Artificial Intelligence*, *6*, 100207. https://doi.org/10.1016/j.caeai.2024.100207

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Funder, D. C. (2001). Accuracy in personality judgment: Research and theory concerning an obvious question. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace.* (pp. 121–140). American Psychological Association. https://doi.org/10.1037/10434-005

Gage, N. L., & Needels, M.C. (1989). Process-product research on teaching: a review of criticisms. *The Elementary School Journal*, 89(3), 253–300.

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

Gold, B., & Windscheid, J. (2020). Observing 360-degree classroom videos – Effects of video type on presence, emotions, workload, classroom observations, and ratings of teaching quality. *Computers & Education, 156*, 103960. https://doi.org/10.1016/j.compedu.2020.103960

Goldinger, S. D., & Papesh, M. H. (2012). Pupil Dilation Reflects the Creation and Retrieval of Memories. *Current Directions in Psychological Science*, *21*(2), 90–95. https://doi.org/10.1177/0963721412436811

Göllner, R., Wagner, W., Eccles, J. S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers? *Journal of Educational Psychology*, *110*(5), 709–725. https://doi.org/10.1037/edu0000236

Göllner, R., Wagner, W., Klieme, E., Lüdtke, O., Nagengast, B., & Trautwein, U. (2016). Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven [Assessing teaching quality with student ratings: Opportunities, limitations and research perspectives]. In Bundesministerium für Bildung und Forschung (Ed.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments* (pp. 63–82). Bundesministerium für Bildung und Forschung. https://doi.org/10.25656/01:12674

Gorman, C. A., Meriac, J. P., Ray, J. L., & Roddy, T. W. (2015). Current trends in rater training: A survey of rater training programs in American organizations. In B. J. O'Leary, B. L. Weathington, C. J. L. Cunningham, & M. D. Biderman (Eds.), *Trends in training* (pp. 1–24). Cambridge Scholars Publishing.

Graham, J. P., Murray, J., & Allen, B. (2023). Immersive learning environments: Do they improve student skills or cause cognitive overload? A literature review. *EDULEARN Proceedings, 15*, 5780–5786. https://doi.org/10.21125/edulearn.2023.1513

Gräsel, C., & Göbel, K. (2011). Unterrichtsqualität [Teaching quality]. In H. Reinders, H. Ditton, C. Gräsel, & B. Gniewosz (Eds.), *Empirische Bildungsforschung* (pp. 87–98). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93021-3_8

Grossman, P. (2021). *Teaching core practices in teacher education*. Harvard Education Press.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English

Language Arts and Teachers' Value-Added Scores. *American Journal of Education*, *119*(3), 445–470. https://doi.org/10.1086/669901

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. https://doi.org/10.25656/01:21187

Hamel, C., & Viau-Guay, A. (2019). Using video to support teachers' reflective practice: A literature review. *Cogent Education*, *6*(1), 1673689. https://doi.org/10.1080/2331186X.2019.1673689

Harvard Graduate School of Education. (2024). *Professional Development. All Programs*. Harvard Graduate School of Education. https://www.gse.harvard.edu/professional-education/programs?aud=All&topic=714&sort_bef_combine=sdate_ASC&page=0%2C%2C0

Hattie, J. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement.* Routledge. https://doi.org/10.4324/9780203887332

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Helmke, A. (2004). *Unterrichtsqualität: Erfassen, Bewerten, Verbessern, 3. Auflage [Teaching quality: Assessing, evaluating, improving, 3rd edition]*. Kallmeyer.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts [Teaching quality and teacher professionalism: diagnosis, evaluation and improvement of teaching].* Klett-Kallmeyer.

Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *The Elementary School Journal*, *101*(1), 3–20.

Hiebert, J., & Stigler, J. W. (2023). Creating Practical Theories of Teaching. In A.-K. Praetorius & C. Y. Charalambous (Eds.), *Theorizing Teaching* (pp. 23–56). Springer International Publishing. https://doi.org/10.1007/978-3-031-25613-4_2

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Bill and Melinda Gates Foundation. Retrieved on November 4, 2024 from https://files.eric.ed.gov/fulltext/ED540957.pdf

Hollins, E. R. (2011). Teacher Preparation For Quality Teaching. *Journal of Teacher Education*, *62*(4), 395–407. https://doi.org/10.1177/0022487111409415

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Hou, R., Fütterer, T., Bühler, B., Bozkir, E., Gerjets, P., Trautwein, U., & Kasneci, E. (2024). Automated Assessment of Encouragement and Warmth in Classrooms Leveraging Multimodal Emotional Features and ChatGPT. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education* (pp. 60–74). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64302-6_5

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*(1), 64–86. https://doi.org/10.1037/1082-989X.5.1.64

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology*, *64*(5), 502–508. https://doi.org/10.1037/0021-9010.64.5.502

Janik, T., & Seidel, T. (2009). *The power of video studies in investigating teaching and learning in the classroom*. Waxmann.

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michaelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. Basic Books.

Jentsch, A., Benecke, K., Blömeke, S., König, J., & Kaiser, G. (2024). Effects of observation mode on ratings of teaching quality in secondary mathematics classrooms. *ZDM*, *56*, 789–800. https://doi.org/10.1007/s11858-024-01557-z

Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of Observation: Considerations for Developing a Classroom Observation System That Helps Districts Achieve Consistent and Accurate Scores. MET Project, Policy and Practice Brief.* Bill

and Melinda Gates Foundation. Retrieved on November 12, 2024 from https://files.eric.ed.gov/fulltext/ED583085.pdf

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Jones, E., & Bergin, C. (2019). Evaluating Teacher Effectiveness Using Classroom Observations: A Rasch Analysis of the Rater Effects of Principals. *Educational Assessment*, *24*(2), 91–118. https://doi.org/10.1080/10627197.2018.1564272

Junker, R., Zucker, V., Oellers, M., Rauterberg, T., Konjer, S., Meschede, N., & Holodinsky, M. (2022). *Lehren und Forschen mit Videos in der Lehrkräftebildung [Teaching and researching with videos in teacher education].* Waxmann.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. https://doi.org/10.1016/0010-0285(76)90015-3

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Bill and Melinda Gates Foundation. Retrieved on November 4, 2024 from http://metproject. org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kavanagh, S., Luxton-Reilly, A., Wüensche, B., & Plimmer, B. (2016). Creating 360° educational video: A case study. *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI '16*, 34–39. https://doi.org/10.1145/3010915.3011001

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. C. (2020). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, *28*(62), 1–34. https://doi.org/10.14507/epaa.28.5012

Keskin, Ö., Seidel, T., Stürmer, K., & Gegenfurtner, A. (2024). Eye-tracking research on teacher professional vision: A meta-analytic review. *Educational Research Review*, *42*, 100586. https://doi.org/10.1016/j.edurev.2023.100586

Kilburn, D. (2014). *Methods for recording video in the classroom: Producing single and multi-camera videos for research into teaching and learning* (NCRM Working Paper). National Centre for Research Methods. http://eprints.ncrm.ac.uk/3599/

Kleickmann, T., Steffensky, M., & Praetorius, A.-K. (2020). Quality of Teaching in Science Education: More Than Three Basic Dimensions? *Zeitschrift für Pädagogik Beiheft*, 66(*1*), 37–53. https://doi.org/10.3262/ZPB2001037

Klette, K. (2016). Introduction: Studying Interaction and Instructional Patterns in Classrooms. In K. Klette, O. K. Bergem, & A. Roe (Eds.), *Teaching and Learning in Lower Secondary Schools in the Era of PISA and TIMSS* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-17302-3_1

Klette, K. (2023). Classroom observation as a means of understanding teaching quality: Towards a shared language of teaching? *Journal of Curriculum Studies*, *55*(1), 49–62. https://doi.org/10.1080/00220272.2023.2172360

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal*, *17*(1), 129–146. https://doi.org/10.1177/1474904117703228

Klieme, E. (2019). Unterrichtsqualität [Teaching quality]. In M. Harring, C. Rohlfs & M. Gläser-Zikuda (Eds.), *Handbuch Schulpädagogik* (pp. 393–408). Waxmann.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik, & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung [Teaching mathematics at lower secondary level: "task culture" and lesson design]. In J. Baumert & E. Klieme (Eds.), *TIMSS – Impulse für Schule und Unterricht, Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 43–58). Bundesministerium für Bildung und Forschung (BMBF).

König, J., Santagata, R., Scheiner, T., Adleff, A.-K., Yang, X., & Kaiser, G. (2022). Teacher noticing: A systematic literature review of conceptualizations, research designs, and findings on learning to notice. *Educational Research Review*, *36*, 100453. https://doi.org/10.1016/j.edurev.2022.100453

Korthagen, F. A. J. (2007). The gap between research and practice revisited. *Educational Research and Evaluation*, *13*(3), 303–310. https://doi.org/10.1080/13803610701640235

Kosel, C., Holzberger, D., & Seidel, T. (2021). Identifying Expert and Novice Visual Scanpath Patterns and Their Relationship to Assessing Learning-Relevant Student Characteristics. *Frontiers in Education*, *5*, 612175. https://doi.org/10.3389/feduc.2020.612175

Kosko, K. W., Ferdig, R. E., Lenart, C., Heisler, J., & Guan, Q. (2024). Exploring teachers' eye-tracking data and professional noticing when viewing a 360 video of elementary mathematics. *Journal of Mathematics Teacher Education, 27*(6), 1–24. https://doi.org/10.1007/s10857-024-09667-x

Kosko, K. W., Ferdig, R. E., & Zolfaghari, M. (2021). Pre-service Teachers' Professional Noticing When Viewing Standard and 360 Video. *Journal of Teacher Education*, *72*(3), 284–297. https://doi.org/10.1177/0022487120939544

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

Krammer, K., & Reusser, K. (2005). Unterrichtsvideos als Medium der Aus- und Weiterbildung von Lehrpersonen [Classroom videos as a tool for teacher training and professional development]. *Beiträge zur Lehrerbildung*, *23*(1), 35–50. https://doi.org/10.36950/bzl.23.1.2005.10146

Kunter, M., & Baumert, J. (2007). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., & Trautwein, U. (2013). *Psychologie des Unterrichts [Psychology of teaching].* Schöningh.

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers. Results from the COACTIV Project* (pp. 97–124). Springer.

Kunz, K., & Zinn, B. (2022). Virtuelle Unterrichtsszenarien in der Lehrpersonenbildung - eine Studie zur Akzeptanz, Immersion und zum Präsenzerleben mit Studierenden der Berufs- und Technikpädagogik [Virtual teaching scenarios in teacher education - a study of acceptance, immersion, and classroom experience with vocational and technical

education students]. *Unterrichtswissenschaft*, 1–25. https://doi.org/10.1007/s42010-022-00151-0

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A Window to the Preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Lawson, A. P., Martella, A. M., LaBonte, K., Delgado, C. Y., Zhao, F., Gluck, J. A., Munns, M. E., Wells LeRoy, A., & Mayer, R. E. (2024). Confounded or Controlled? A Systematic Review of Media Comparison Studies Involving Immersive Virtual Reality for STEM Education. *Educational Psychology Review*, *36*(3), 69–104. https://doi.org/10.1007/s10648-024-09908-8

Lehikko, A., Nykänen, M., Lukander, K., Uusitalo, J., & Ruokamo, H. (2024). Exploring interactivity effects on learners' sense of agency, cognitive load, and learning outcomes in immersive virtual reality: A mixed methods study. *Computers & Education: X Reality*, *4*, 100066. https://doi.org/10.1016/j.cexr.2024.100066

Lehrkräftefortbildung Baden-Württemberg. (2024). *LFB Online-Suche Veranstaltungstermine.* Lehrkräftefortbildung Baden-Württemberg. https://lfbo.kultus-bw.de/lfb/suche

Lenske, G. (2016). *Schülerfeedback in der Grundschule: Untersuchung zur Validität [Student feedback in elementary school: investigation of validity]*. Waxmann.

Letzring, T. D., & Funder, D. C. (2019). The Realistic Accuracy Model. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment* (pp. 9–22). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.2

Li, H., Liu, J., & Hunter, C. V. (2020). A Meta-Analysis of the Factor Structure of the Classroom Assessment Scoring System (CLASS). *The Journal of Experimental Education*, *88*(2), 265–287. https://doi.org/10.1080/00220973.2018.1551184

Li, W., Feng, Q., Zhu, X., Yu, Q., & Wang, Q. (2023). Effect of summarizing scaffolding and textual cues on learning performance, mental model, and cognitive load in a virtual reality environment: An experimental study. *Computers & Education*, *200*, Article 104793. https://doi.org/10.1016/j.compedu.2023.104793

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of

the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537. https://doi.org/10.1016/j.learninstruc.2008.11.001

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61–95. https://doi.org/10.1007/s11092-018-09291-3

Lotz, M., Gabriel, K., & Lipowsky, F. (2013). Niedrig und hoch inferente Verfahren der Unterrichtsbeobachtung. Analysen zu deren gegenseitiger Validierung [Low and high inferential methods of classroom observation. Analyses of their mutual validation]. *Zeitschrift für Pädagogik*, *59*(3), 357–380. https://doi.org/10.25656/01:11942

Marder, M., & Walkington, C. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measuring Effective Teaching project* (pp. 234–277). Wiley.

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation*, *49*, 15–29. https://doi.org/10.1016/j.stueduc.2016.03.002

Martin-Raugh, M., Tannenbaum, R. J., Tocci, C. M., & Reese, C. (2016). Behaviorally anchored rating scales: An application for evaluating teaching practice. *Teaching and Teacher Education*, *59*, 414–419. https://doi.org/10.1016/j.tate.2016.07.026

McGarr, O., O'Grady, E., & Guilfoyle, L. (2017). Exploring the theory-practice gap in initial teacher education: Moving beyond questions of relevance to issues of power and authority. *Journal of Education for Teaching*, *43*(1), 48–60. https://doi.org/10.1080/02607476.2017.1256040

Meyer, H. (2003). Zehn Merkmale guten Unterrichts. Empirische Befunde und didaktische Ratschläge [Ten characteristics of good teaching. Empirical findings and didactic advice.]. *Pädagogik, 55*(10), 36–43.

Moore, T. M., & Picou, E. M. (2018). A potential bias in subjective ratings of mental effort. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2405-2421. https://doi.org/10.1044/2018_JSLHR-H-17-0451

Muñiz-Rodríguez, L., Alonso, P., Rodríguez-Muñiz, L. J., De Coninck, K., Vanderlinde, R., & Valcke, M. (2018). Exploring the Effectiveness of Video-Vignettes to Develop Mathematics Student Teachers' Feedback Competence. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(11). https://doi.org/10.29333/ejmste/92022

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Noroozi, O., Pijeira-Díaz, H. J., Sobocinski, M., Dindar, M., Järvelä, S., & Kirschner, P. A. (2020). Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review. *Education and Information Technologies*, *25*(6), 5499–5547. https://doi.org/10.1007/s10639-020-10229-w

O'Leary, M. (2020). *Classroom Observation: A Guide to the Effective Observation of Teaching and Learning (2nd ed.)*. Routledge. https://doi.org/10.4324/9781315630243

Otrel-Cass, K., Cowie, B., & Maguire, M. (2010). Taking video cameras into the classroom. *Waikato Journal of Education*, *15*(2), 109–118. https://doi.org/10.15663/wje.v15i2.117

Panayiotou, A., Herbert, B., Sammons, P., & Kyriakides, L. (2021). Conceptualizing and exploring the quality of teaching using generic frameworks: A way forward. *Studies in Educational Evaluation*, *70*, 101028. https://doi.org/10.1016/j.stueduc.2021.101028

Parong, J., & Mayer, R. E. (2021). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning*, *37*(1), 226–241. https://doi.org/10.1111/jcal.12482

Paulicke, P., Ehmke, T., Pietsch, M., & Schmidt, T. (2019). Wie beeinflusst die Kameraperspektive die Beurteilung der Unterrichtsqualität? [How does the camera perspective influence the assessment of teaching quality?]. *Zeitschrift für Bildungsforschung*, *9*(3), 411–435. https://doi.org/10.1007/s35834-019-00246-2

Petko, D., Waldis, M., Pauli, C., & Reusser, K. (2003). Methodologische Überlegungen zur videogestützten Forschung in der Mathematikdidaktik: Ansätze der TIMSS 1999 Video Studie und ihrer schweizerischen Erweiterung [Methodological considerations on video-based research in mathematics didactics: approaches of the TIMSS 1999 video

study and its Swiss extension]. *Zentralblatt für Didaktik der Mathematik*, *35*(6), 265–280. https://doi.org/10.1007/BF02656691

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3.* Paul H. Brookes Publishing Co.

Piwowar, V., Barth, V. L., Ophardt, D., & Thiel, F. (2018). 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. (2014). *Messung von Unterrichtsqualität durch Ratings [Assessing teaching quality through ratings].* Waxmann.

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A.-K., & Charalambous, C. Y. (2023). Where Are We on Theorizing Teaching? A Literature Overview. In A.-K. Praetorius & C. Y. Charalambous (Eds.), *Theorizing Teaching* (pp. 1–22). Springer International Publishing. https://doi.org/10.1007/978-3-031-25613-4_1

Praetorius, A. K., Grünkorn, J., & Klieme, E. (2020). Towards developing a theory of generic teaching quality: Origin, current status, and necessary next steps regarding the three basic dimensions model. *Zeitschrift für Pädagogik Beiheft, 66*(1), 15–36. https://doi.org/10.1007/ZPB2001015

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Prilop, C. N., Weber, K. E., & Kleinknecht, M. (2020). Effects of digital video-based feedback environments on pre-service teachers' feedback competence. *Computers in Human Behavior*, *102*, 120–131. https://doi.org/10.1016/j.chb.2019.08.011

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern [Motivational support in mathematics lessons: Teaching from the perspective of learners and observers]*. Waxmann.

Ranieri, M., Luzzi, D., Cuomo, S., & Bruni, I. (2022). If and how do 360° videos fit into education settings? Results from a scoping review of empirical research. *Journal of Computer Assisted Learning*, *38*(5), 1199–1219. https://doi.org/10.1111/jcal.12683

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Reusser, K. (2008). *Lernwirksamer Unterricht—Das Kerngeschäft von Lehrpersonen [Effective teaching - the core business of teachers]*. https://doi.org/10.5167/UZH-14588

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Roche, L., Kittel, A., Cunningham, I., & Rolland, C. (2021). 360° Video Integration in Teacher Education: A SWOT Analysis. *Frontiers in Education*, *6*, 761176. https://doi.org/10.3389/feduc.2021.761176

Rodriguez, L. A., Swain, W. A., & Springer, M. G. (2020). Sorting Through Performance Evaluations: The Influence of Performance Evaluation Reform on Teacher Attrition and Mobility. *American Educational Research Journal*, *57*(6), 2339–2377. https://doi.org/10.3102/0002831220910989

Rømer, T. A. (2019). A critique of John Hattie's theory of Visible Learning. *Educational Philosophy and Theory*, *51*(6), 587–598. https://doi.org/10.1080/00131857.2018.1488216

Rosendahl, P., & Wagner, I. (2023). 360° videos in education – A systematic literature review on application areas and future potentials. *Education and Information Technologies*. https://doi.org/10.1007/s10639-022-11549-9

Rupp, M. A., Odette, K. L., Kozachuk, J., Michaelis, J. R., Smither, J. A., & McConnell, D. S. (2019). Investigating learning outcomes and subjective experiences in 360-degree videos. *Computers & Education*, *128*, 256–268. https://doi.org/10.1016/j.compedu.2018.09.015

Ruth-Herbein, E., Maier, J. L., & Fauth, B. (2022). Promoting Teaching Quality Through Classroom Observation and Feedback: Design of a Program in the German State of Baden-Württemberg. In J. Manzi, Y. Sun, & M. R. García (Eds.), *Teacher Evaluation Around the World* (pp. 271–289). Springer International Publishing. https://doi.org/10.1007/978-3-031-13639-9_12

Santagata, R., Gallimore, R., & Stigler, J. W. (2005). The use of video teaching for teacher education and professional development. In C. Vrasidas & G. V. Glass (Eds.), *Preparing teachers to teach with technology: Current perspectives on applied information technologies* (pp. 151–167). Information Age Publishing.

Santagata, R., König, J., Scheiner, T., Nguyen, H., Adleff, A.-K., Yang, X., & Kaiser, G. (2021). Mathematics teacher learning to notice: A systematic review of studies of video-based programs. *ZDM*, *53*(1), 119–134. https://doi.org/10.1007/s11858-020-01216-z

Sauerwein, M., & Klieme, E. (2016). Anmerkungen zum Qualitätsbegriff in der Bildungsforschung [Notes on the concept of quality in educational research]. *Swiss Journal of Educational Research*, *38*(3), 459–478. https://doi.org/10.24452/sjer.38.3.4988

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Pergamon.

Scheerens, J., Luyten, J. W., Steen, R., & de Thouars, Y. C. H. (2007). *Review and meta-analyses of school and teaching effectiveness*. Universiteit Twente.

Scheiter, K. (2021). Lernen und Lehren mit digitalen Medien: Eine Standortbestimmung [Learning and teaching with digital media: An assessment of the current situation]. *Zeitschrift für Erziehungswissenschaft*, *24*(5), 1039–1060. https://doi.org/10.1007/s11618-021-01047-y

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality. *Frontiers in Psychology, 7,* Article 1105. https://doi.org/10.3389/fpsyg.2016.00110

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, *48*(1–2), 29–40. https://doi.org/10.1007/s11858-016-0765-0

Sigurjónsson, J. Ö., Sigurðardóttir, A. K., Gísladóttir, B., & van Bommel, J. (2022). Connecting student perceptions and classroom observations as measures of cognitive activation. *Nordic Studies in Education*, *42*(4), 328–346. https://doi.org/10.23865/nse.v42.3636

Seidel, T. (2022). Einleitung - Videobasierte Forschung und ihr Beitrag zu einer verbesserten Lehrkräftebildung in Deutschland [Introduction - Video-based research and its contribution to improved teacher training in Germany]. In R. Junker, V. Zucker, M. Oellers, T. Rauterberg, S. Konjer, N. Meschede, & M. Holodynski, (Eds.), *Lehren und Forschen mit Videos in der Lehrkräftebildung.* (pp. 2–4). Waxmann.

Seidel, T., Farrell, M., Martin, M., Rieß, W., & Renkl, A. (2022). Developing scripted video cases for teacher education: Creating evidence-based practice representations using mock ups. *Frontiers in Education*, *7*, 965498. https://doi.org/10.3389/feduc.2022.965498

Seidel, T., & Prenzel, M. (2010). Beobachtungsverfahren: Vom Datenmaterial zur Datenanalyse [Observation methods: From data set to data analysis]. In H. Holling & B. Schmitz (Eds.), *Handbuch Statistik, Methoden und Evaluation* (pp. 139–152). Hogrefe.

Seidel, T., & Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research*, *77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Seidel, T., & Stürmer, K. (2014). Modeling and Measuring the Structure of Professional Vision in Pre-service Teachers. *American Educational Research Journal*, *51*(4), 739–771. https://doi.org/10.3102/0002831214531321

Seidel, T., & Thiel, F. (2017). Standards und Trends der videobasierten Lehr-Lernforschung [Standards and trends in video-based teaching and learning research]. *Zeitschrift für Erziehungswissenschaft, 20*(1), 1–21. https://doi.org/10.1007/s11618-017-0726-6

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

Senden, B., Nilsen, T., & Blömeke, S. (2022). 5. Instructional Quality: A Review of Conceptualizations, Measurement Approaches, and Research Findings. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of Analyzing Teaching Quality* (pp. 140–172). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021-05

Shek, M. M.-P., Leung, K.-C., & To, P. Y.-L. (2021). Using a video annotation tool to enhance student-teachers' reflective practices and communication competence in consultation practices through a collaborative learning community. *Education and Information Technologies*, *26*(4), 4329–4352. https://doi.org/10.1007/s10639-021-10480-9

Sherin, M., & van Es, E. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education*, *13*(3), 475–491.

Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, *6*(6), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Snelson, C., & Hsu, Y.-C. (2020). Educational 360-Degree Videos in Virtual Reality: A Scoping Review of the Emerging Research. *TechTrends*, *64*(3), 404–412. https://doi.org/10.1007/s11528-019-00474-3

Storms, M. D. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality and Social Psychology*, *27*(2), 165–175. https://doi.org/10.1037/h0034782

Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2023). Multimodal Engagement Analysis From Facial Videos in the Classroom. *IEEE Transactions on Affective Computing*, *14*(2), 1012–1027. https://doi.org/10.1109/TAFFC.2021.3127692

Sweller, J. (2011). *Cognitive Load Theory*. Springer. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Syring, M., Bohl, T., Kleinknecht, M., Kuntze, S., Rehm, M., & Schneider, J. (2015). Video or text in case-based teacher education? An examination of the effects of different media on cognitive load and motivational-emotional processes in case-based learning. *Zeitschrift für Erziehungswissenschaft*, *18*, 667–685. https://doi.org/10.1007/s11618-015-0631-9

Tarkian, J., Lankes, E. M., & Thiel, F. (2019). Externe Evaluation - Konzeption und Implementation in den 16 Ländern [External evaluation - conceptualization and implementation in the 16 federal states]. In F. Thiel, J. Tarkian, E.-M. Lankes, N. Maritzen, T. Rieke-Baulecke & A. Kroupa (Eds.), *Datenbasierte Qualitätssicherung und -entwicklung in Schulen* (pp. 105–183). Springer. https://doi.org/10.1007/978-3-658-23240-5_5

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, *46*, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003

Tengberg, M., van Bommel, J., Nilsberth, M., Walkert, M., & Nissen, A. (2022). The quality of instruction in Swedish lower secondary language arts and mathematics. *Scandinavian Journal of Educational Research*, *66*(5), 760–777. https://doi.org/10.1080/00313831.2021.1910564

Tripp, T., & Rich, P. (2012). Using video to analyze one's own teaching. *British Journal of Educational Technology*, *43*(4), 678–704. https://doi.org/10.1111/j.1467-8535.2011.01234.x

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology*, *93*(3), 711–719. https://doi.org/10.1037/0021-9010.93.3.711

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.

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

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., Depaepe, F., & Van Den Noortgate, W. (2021). Towards measuring cognitive load through

multimodal physiological data. *Cognition, Technology & Work*, *23*(3), 567–585. https://doi.org/10.1007/s10111-020-00641-0

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*(2), 281–300. https://doi.org/10.1037/a0017908

Vieluf, S., & Klieme, E. (2023). Teaching Effectiveness Revisited Through the Lens of Practice Theories. In A.-K. Praetorius & C. Y. Charalambous (Eds.), *Theorizing Teaching* (pp. 57–95). Springer International Publishing. https://doi.org/10.1007/978-3-031-25613-4_3

Vinokic, K., Begrich, L., Kunter, M., & Kuger, S. (2024). The Underlying Cognitive Processes of Thin Slices Judgments on Teaching Quality. *Frontline Learning Research*, *12*(3), 69–98. https://doi.org/10.14786/flr.v12i3.1421

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721. https://doi.org/10.1037/edu0000075

Walberg, H. J. (1981). A psychological theory of educational productivity. *Institution of Education Sciences*. https://eric.ed.gov/?id=ED206042

Walshe, N., & Driver, P. (2019). Developing reflective trainee teacher practice with 360-degree video. *Teaching and Teacher Education*, *78*, 97–105. https://doi.org/10.1016/j.tate.2018.11.009

Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., & Munoz, D. P. (2018). Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional Face Task. *Frontiers in Neurology*, *9*, 1029. https://doi.org/10.3389/fneur.2018.01029

Wang, J., & Engelhard, G. (2019). Conceptualizing Rater Judgments and Rating Processes for Rater-Mediated Assessments. *Journal of Educational Measurement*, *56*(3), 582–609. https://doi.org/10.1111/jedm.12226

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*(3), 249–294. https://doi.org/10.3102/00346543063003249

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

White, M. (2022). 3. A Validity Framework for the Design and Analysis of Studies Using Standardized Observation Systems. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of Analyzing Teaching Quality* (pp. 89–120). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021-03

White, M. C. (2018). Rater Performance Standards for Classroom Observation Instruments. *Educational Researcher*, *47*(8), 492–501. https://doi.org/10.3102/0013189X18785623

White, M., & Klette, K. (2023). What's in a score? Problematizing interpretations of observation scores. *Studies in Educational Evaluation*, *77*, 101238. https://doi.org/10.1016/j.stueduc.2023.101238

White, M., & Klette, K. (2024). Signal, error, or bias? Exploring the uses of scores from observation systems. *Educational Assessment, Evaluation and Accountability, 36*, 505–528. https://doi.org/10.1007/s11092-024-09427-8

White, M., & Ronfeldt, M. (2024). Monitoring Rater Quality in Observational Systems: Issues Due to Unreliable Estimates of Rater Quality. *Educational Assessment*, *29*(2), 124–146. https://doi.org/10.1080/10627197.2024.2354311

Wilkes, T., Stark, L., Trempler, K., & Stark, R. (2022). Contrastive Video Examples in Teacher Education: A Matter of Sequence and Prompts. *Frontiers in Education*, *7*, 869664. https://doi.org/10.3389/feduc.2022.869664

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wyss, C., Bäuerlein, K., & Mahler, S. (2023). Pre-service and in-service teachers' professional vision depending on the video perspective—What teacher gaze and verbal reports can tell us. *Frontiers in Education*, *8*, 1282992. https://doi.org/10.3389/feduc.2023.1282992

# 8

# APPENDIX

# 8 Appendix

As part of a collaborative research project involving the Institute for Educational Analysis Baden-Württemberg (IBBW), the Hector Research Institute of Education Sciences and Psychology in Tübingen, and the Universities of Education Freiburg and Heidelberg, a series of staged mathematics classroom videos was developed throughout this dissertation. Five of these videos were employed in a lab study, where pre-service teachers observed the lessons and evaluated teaching quality using the Unterrichtsfeedbackbogen Tiefenstrukturen (UFB), a standardized classroom observation instrument. The following section provides the original German scripts for these five classroom videos, which served as the foundation for data collection in two studies included in this dissertation.

| Video | Creators of the script | Creators of the video |
|---|---|---|
| Distributive law: Positive example *focus on key concepts*, negative example *classroom disruptions* | Anika Dreher, Ann-Kathrin Jaekel, Benjamin Fauth, Evelin Ruth-Herbein, Linn Hansen, Marita Friesen, Richard Göllner, Timo Leuders, Tosca Daltoè | Anika Dreher, Ann-Kathrin Jaekel, Benjamin Fauth, Evelin Ruth-Herbein, Linn Hansen, Marita Friesen, Richard Göllner, Timo Leuders, Tosca Daltoè, Ulrich Trautwein |
| Distributive law: Positive example *focus on key concepts*, positive example *classroom disruptions* | Anika Dreher, Linn Hansen, Marita Friesen, Timo Leuders | Anika Dreher, Ann-Kathrin Jaekel, Benjamin Fauth, Evelin Ruth-Herbein, Linn Hansen, Marita Friesen, Richard Göllner, Timo Leuders, Tosca Daltoè, Ulrich Trautwein |
| Distributive law: Negative example *focus on key concepts*, positive example *classroom disruptions* | Anika Dreher, Linn Hansen, Marita Friesen, Timo Leuders | Anika Dreher, Ann-Kathrin Jaekel, Benjamin Fauth, Evelin Ruth-Herbein, Linn Hansen, Marita Friesen, Richard Göllner, Timo Leuders, Tosca Daltoè, Ulrich Trautwein |
| Calculation of the area of a circle sector: Positive example *challenging tasks* | Anika Dreher, Linn Hansen, Marita Friesen, Dagmar Fischer | Anika Dreher, Ann-Kathrin Jaekel, Benjamin Fauth, Dagmar Fischer, Evelin Ruth-Herbein, Linn Hansen, Marita Friesen, Richard Göllner, Tosca Daltoè, Ulrich Trautwein |
| Calculation of the area of a circle sector: Negative example *challenging tasks* | Anika Dreher, Linn Hansen, Marita Friesen, Dagmar Fischer | Anika Dreher, Ann-Kathrin Jaekel, Benjamin Fauth, Dagmar Fischer, Evelin Ruth-Herbein, Linn Hansen, Marita Friesen, Richard Göllner, Tosca Daltoè, Ulrich Trautwein |

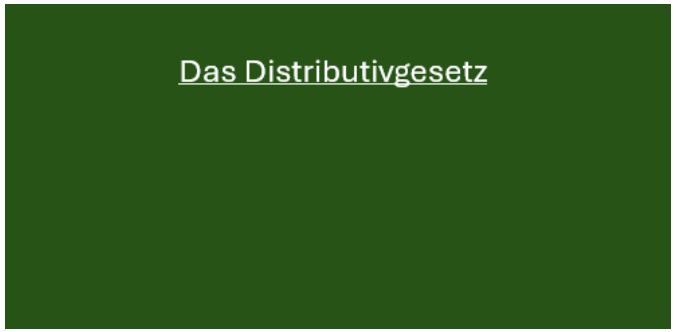
**Script for Video 1: Distributive law**

Positive example *focus on key concepts*

Negative example *classroom disruptions*

LEHRKRAFT
So, wir haben uns letzte Stunde ja schon damit beschäftigt, wie man Terme vereinfachen kann und welche Rechengesetze es dafür gibt. Was wir uns noch nicht überlegt haben, ist, wie man bei der Multiplikation Klammern auflösen und einbauen kann. Die Distributivgesetze, die fehlen uns für das Rechnen mit Variablen noch.

*Die Lehrkraft schreibt „Das Distributivgesetz" an die Tafel.*
*Eine Gruppe an Schülerinnen und Schülern tuschelt hörbar, beim Umdrehen schaut die Lehrkraft die Gruppe ermahnend an.*




Wisst ihr denn noch, was das Distributivgesetz für die Multiplikation sagt?

*Es melden sich Sophia, Levin und Ana.*

Sophia!

SOPHIA
War das das mit dem Ausklammern und Ausmultiplizieren?

LEHRKRAFT
Ja genau. Wer kann denn da mal ein Beispiel dazu machen?

*Die Lehrkraft wartet 15 Sekunden, weil sich niemand meldet.*

Wir hatten das beim geschickten Rechnen zum Beispiel genutzt, fällt euch dazu was ein?

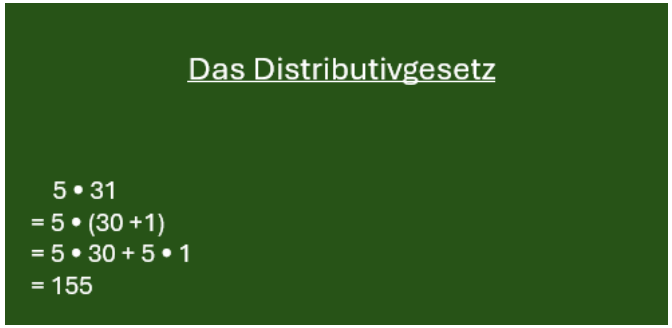
*Es meldet sich nur Maja.*

Maja?

MAJA
War das sowas wie 5•31, wo man dann die 31 in 30 und 1 zerlegt und dann einfach 5•30 und 5•1 rechnet?

LEHRKRAFT
Sehr gut. Magst du das mal an die Tafel schreiben, damit sich alle wieder dran erinnern?

*Maja geht vor an die Tafel und schreibt: 5•31=5•(30+1) = 5•30+5•1=155*
*Zwei Schülerinnen tauschen Zettel hin und her, Lautstärkepegel steigt, Lehrerin ermahnt kurz zur Ruhe.*




MAJA
So?

LEHRKRAFT
Ja. Wo genau hat Maja denn jetzt das Distributivgesetz genutzt?

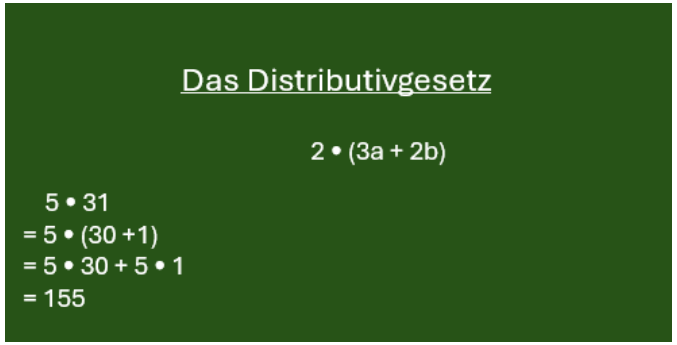
*Es melden sich 2 Hannahs und Klara.*

Klara?

KLARA
Naja, da wo sie die Klammer aufgelöst hat.

LEHRKRAFT
Ganz genau. Wenn wir jetzt mit Variablen rechnen wollen, ist dieses Distributivgesetz besonders wichtig, weil wir dann manchmal gar keine Möglichkeit haben, die Klammern sonst auszurechnen.
Wir machen mal ein Beispiel:
Stellt euch vor, wir haben den Term 2•(3a+2b).

*Die Lehrkraft schreibt den Term an die Tafel.*




Und den wollen wir vereinfachen.
Was ist denn da das Problem?

*Es melden sich Ana, Theo und Tim.*
*Die Schülerinnen und Schüler sind beim Melden ungeduldig und laut.*

Tim?

TIM
Wir wissen halt nicht wie viel 3a+2b ist.

LEHRKRAFT
Genau, das ist das Problem. a und b sind ja Variablen.
Ihr könnt euch das als Streckenlängen vorstellen. Die nennen wir a und b, wenn wir nicht wissen, wie lang sie sind und wenn sie sich verändern können. Machen wir uns mal ein Bild dazu.
Der Term beschreibt hier den Flächeninhalt eines Rechtecks: Eine Seitenlänge ist 2 und die andere Seitenlänge ist 3-mal die Länge a, also zum Beispiel so lang und 2-mal die Länge b, das könnte zum Beispiel so lang sein.

*Die Lehrkraft zeichnet während des Sprechens die Seitenlängen an die Tafel und ergänzt dann die Unterteilungen.*

Habt ihr jetzt eine Idee, wie man diesen Flächeninhalt noch anders beschreiben könnte?

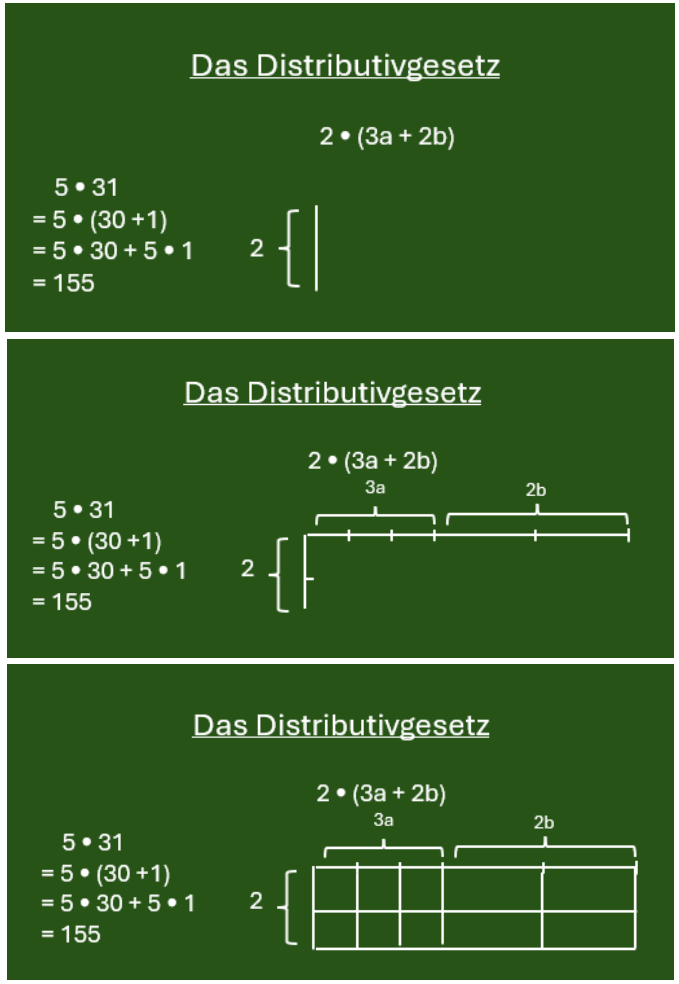
*Es melden sich Steffen, Levin, Sophia und Maja.*

Steffen?

STEFFEN
Naja, das ist 6-mal das eine Rechteck, also das mit dem Flächeninhalt 1•a und dann noch 4-mal das andere Rechteck mit dem Flächeninhalt 1•b.

LEHRKRAFT
Aha. Ich kann denselben Flächeninhalt also auch anders beschreiben: 6 Rechtecke mit Flächeninhalt a und 4 Rechtecke mit Flächeninhalt b.

*Die Lehrkraft zeigt während des Sprechens auf die Rechtecke an der Tafel.*



Wie können wir den Term also vereinfachen?

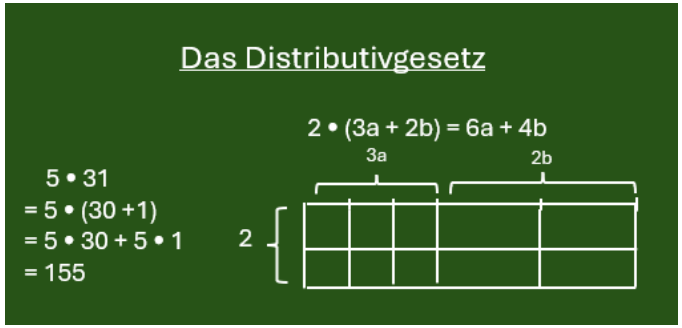
*Es melden sich Steffen, Sara, Levin und Maja.*

*Sara?*

*SARA*
*6a+4b.*

*LEHRKRAFT*
Sehr gut.

*Die Lehrkraft schreibt an die Tafel und vervollständigt die Gleichung.*
*Während die Lehrkraft an die Tafel schreibt, machen zwei Schüler Faxen hinter ihrem Rücken.*




Das heißt, wir können den Term 2•(3a+2b) vereinfachen, indem wir 6a+4b draus machen, weil beide Terme denselben Flächeninhalt beschreiben. Schaut euch mal an, was wir da gemacht haben: 2•3a und 2•2b.
Wir multiplizieren den Faktor vor der Klammer mit den beiden Summanden in der Klammer.

*Die Lehrkraft zeigt in der Gleichung an der Tafel während des Sprechens.*

Genau wie ohne Variablen. Das Muster ist das gleiche.
Nur dass wir mit Variablen aufpassen müssen, dass wir nicht sagen können, welchen Wert sie haben, weil sie sich verändern können.

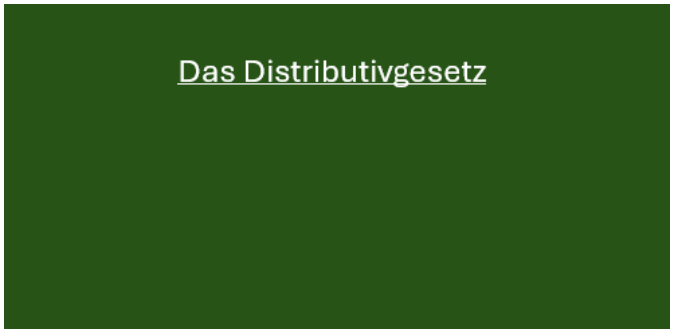
**Script for Video 2: Distributive law**

Positive example *focus on key concepts*

Positive example *classroom disruptions*

LEHRKRAFT
So, wir haben uns letzte Stunde ja schon damit beschäftigt, wie man Terme vereinfachen kann und welche Rechengesetze es dafür gibt. Was wir uns noch nicht überlegt haben, ist, wie man bei der Multiplikation Klammern auflösen und einbauen kann. Die Distributivgesetze, die fehlen uns für das Rechnen mit Variablen noch.

*Die Lehrkraft schreibt „Das Distributivgesetz" an die Tafel.*




Wisst ihr denn noch, was das Distributivgesetz für die Multiplikation sagt?

*Es melden sich Sophia, Levin und Ana.*

Sophia!

SOPHIA
War das das mit dem Ausklammern und Ausmultiplizieren?

LEHRKRAFT
Ja genau. Wer kann denn da mal ein Beispiel dazu machen?

*Die Lehrkraft wartet 15 Sekunden, weil sich niemand meldet.*

Wir hatten das beim geschickten Rechnen zum Beispiel genutzt, fällt euch dazu was ein?

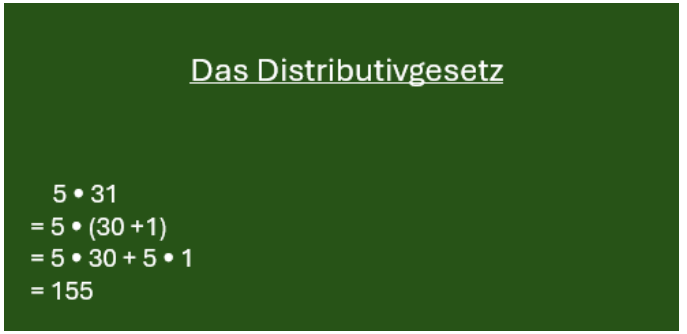
*Es meldet sich nur Maja.*

Maja?

MAJA
War das sowas wie 5•31, wo man dann die 31 in 30 und 1 zerlegt und dann einfach 5•30 und 5•1 rechnet?

LEHRKRAFT
Sehr gut. Magst du das mal an die Tafel schreiben, damit sich alle wieder dran erinnern?

*Maja geht vor an die Tafel und schreibt: 5•31=5•(30+1) = 5•30+5•1=155*




MAJA
So?

LEHRKRAFT
Ja. Wo genau hat Maja denn jetzt das Distributivgesetz genutzt?

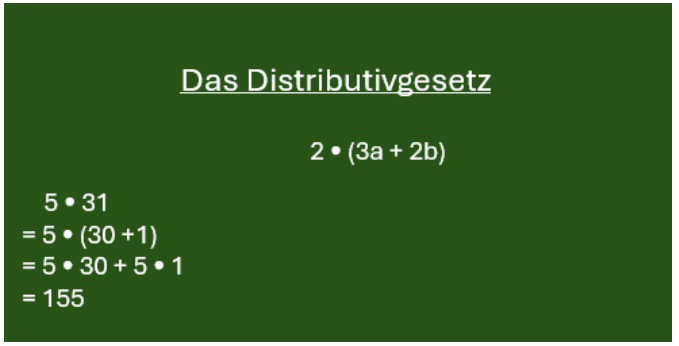
*Es melden sich 2 Hannahs und Klara.*

Klara?

KLARA
Naja, da wo sie die Klammer aufgelöst hat.

LEHRKRAFT
Ganz genau. Wenn wir jetzt mit Variablen rechnen wollen, ist dieses Distributivgesetz besonders wichtig, weil wir dann manchmal gar keine Möglichkeit haben, die Klammern sonst auszurechnen.
Wir machen mal ein Beispiel:
Stellt euch vor, wir haben den Term 2•(3a+2b)

*Die Lehrkraft schreibt den Term an die Tafel.*




Und den wollen wir vereinfachen.
Was ist denn da das Problem?

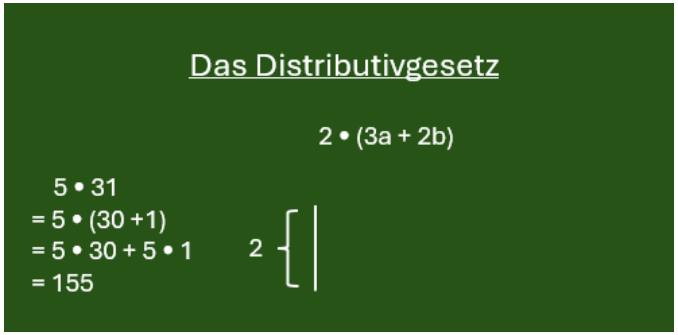
*Es melden sich Ana, Theo und Tim.*

Tim?

TIM
Wir wissen halt nicht wie viel 3a+2b ist.

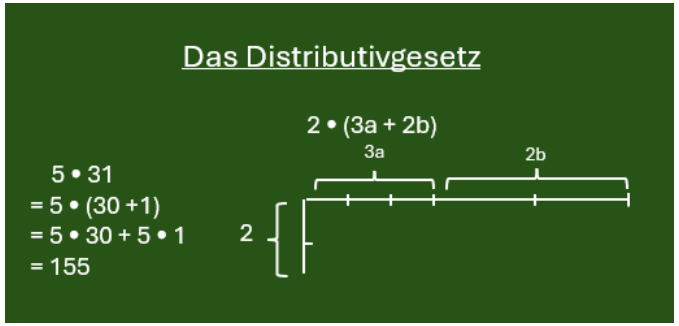
LEHRKRAFT
Genau, das ist das Problem. a und b sind ja Variablen.
Ihr könnt euch das als Streckenlängen vorstellen. Die nennen wir a und b, wenn wir nicht wissen, wie lang sie sind und wenn sie sich verändern können. Machen wir uns mal ein Bild dazu.
Der Term beschreibt hier den Flächeninhalt eines Rechtecks: Eine Seitenlänge ist 2 und die andere Seitenlänge ist 3-mal die Länge a, also zum Beispiel so lang und 2-mal die Länge b, das könnte zum Beispiel so lang sein.

*Die Lehrkraft zeichnet während des Sprechens die Seitenlängen an die Tafel und ergänzt dann die Unterteilungen.*




Habt ihr jetzt eine Idee, wie man diesen Flächeninhalt noch anders beschreiben könnte?

*Es melden sich Steffen, Levin, Sophia und Maja.*

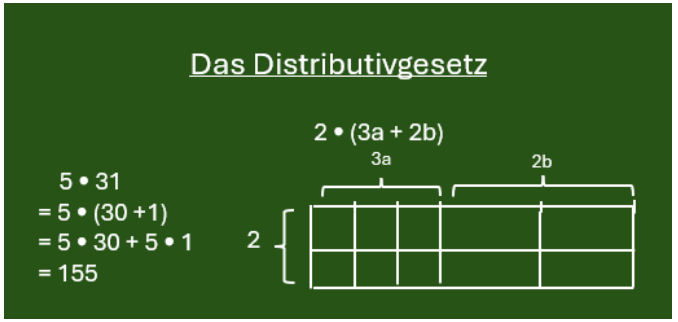



Steffen?

STEFFEN
Naja, das ist 6-mal das eine Rechteck, also das mit dem Flächeninhalt 1•a und dann noch 4-mal das andere Rechteck mit dem Flächeninhalt 1•b.

LEHRKRAFT
Aha. Ich kann denselben Flächeninhalt also auch anders beschreiben: 6 Rechtecke mit Flächeninhalt a und 4 Rechtecke mit Flächeninhalt b.

*Die Lehrkraft zeigt während des Sprechens auf die Rechtecke an der Tafel.*




Wie können wir den Term also vereinfachen?

*Es melden sich Steffen, Sara, Levin und Maja.*

Sara?

SARA
6a+4b.

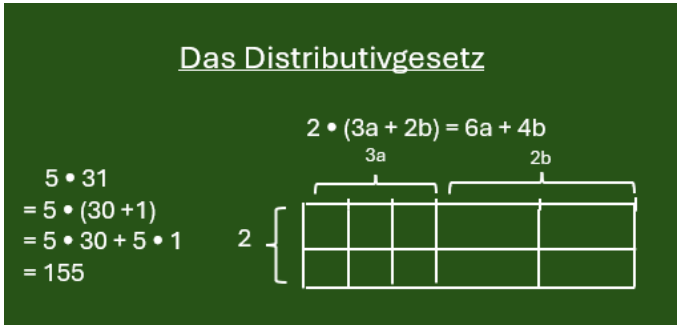
LEHRKRAFT
Sehr gut.

*Die Lehrkraft schreibt an die Tafel und vervollständigt die Gleichung.*

Das heißt, wir können den Term 2•(3a+2b) vereinfachen, indem wir 6a+4b draus machen, weil beide Terme denselben Flächeninhalt beschreiben. Schaut euch mal an, was wir da gemacht haben: 2•3a und 2•2b.
Wir multiplizieren den Faktor vor der Klammer mit den beiden Summanden in der Klammer.

*Die Lehrkraft zeigt in der Gleichung an der Tafel während des Sprechens.*

Genau wie ohne Variablen. Das Muster ist das gleiche.
Nur dass wir mit Variablen aufpassen müssen, dass wir nicht sagen können, welchen Wert sie haben, weil sie sich verändern können.



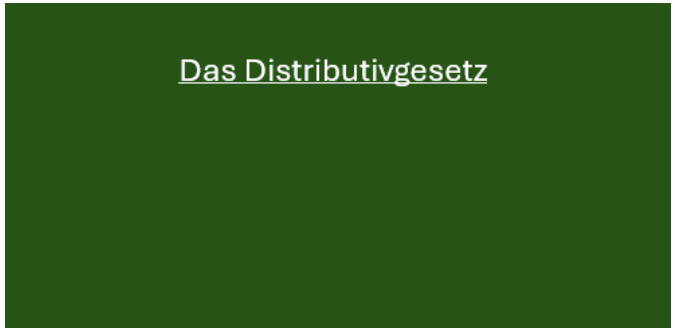
**Script for Video 3: Distributive law**

Negative example *focus on key concepts*

Positive example *classroom disruptions*

LEHRKRAFT
So, wir haben uns letzte Stunde ja schon damit beschäftigt, wie man Terme vereinfachen kann und welche Rechengesetze es dafür gibt. Was wir uns noch nicht überlegt haben, ist, wie man bei der Multiplikation Klammern auflösen und einbauen kann. Die Distributivgesetze, die fehlen uns für das Rechnen mit Variablen noch.

*Die Lehrkraft schreibt „Das Distributivgesetz" an die Tafel.*




Wisst ihr denn noch, was das Distributivgesetz für die Multiplikation sagt?

*Es melden sich Sophia, Levin und Ana.*

Sophia!

SOPHIA
War das das mit dem Ausklammern und Ausmultiplizieren?

LEHRKRAFT
Ja genau. Wer kann denn da mal ein Beispiel dazu machen?

*Die Lehrkraft wartet 15 Sekunden, weil sich niemand meldet.*

Wir hatten das beim geschickten Rechnen zum Beispiel genutzt, fällt euch dazu was ein?

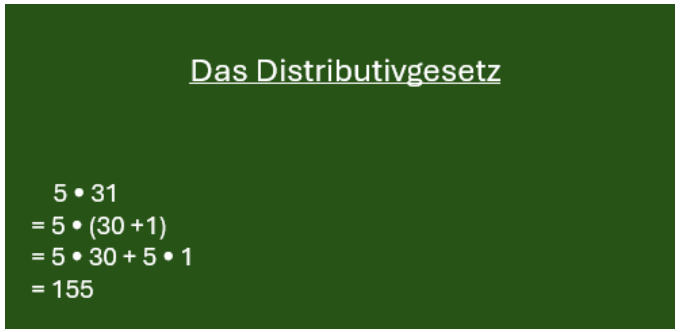
*Es meldet sich nur Maja.*

Maja?

MAJA
War das sowas wie 5•31, wo man dann die 31 in 30 und 1 zerlegt und dann einfach 5•30 und 5•1 rechnet?

LEHRKRAFT
Sehr gut. Magst du das mal an die Tafel schreiben, damit sich alle wieder dran erinnern?

*Maja geht vor an die Tafel und schreibt: 5•31=5•(30+1) = 5•30+5•1=155*




MAJA
So?

LEHRKRAFT
Ja. Wo genau hat Maja denn jetzt das Distributivgesetz genutzt?

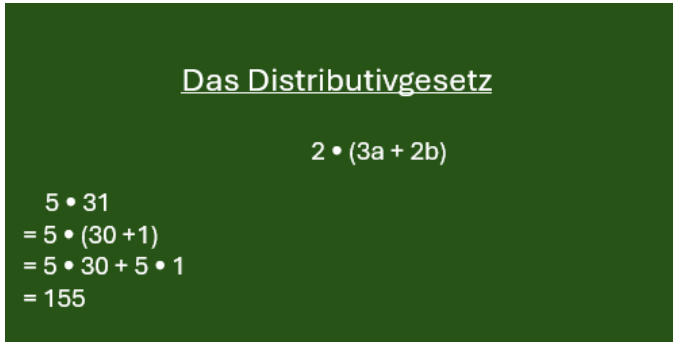
*Es melden sich 2 Hannahs und Klara.*

Klara?

KLARA
Naja, da wo sie die Klammer aufgelöst hat.

LEHRKRAFT
Ganz genau. Wenn wir jetzt mit Variablen rechnen wollen, ist dieses Distributivgesetz besonders wichtig, weil wir dann manchmal gar keine Möglichkeit haben, die Klammern sonst auszurechnen.
Wir machen mal ein Beispiel:
Stellt euch vor, wir haben den Term 2•(3a+2b)

*Die Lehrkraft schreibt den Term an die Tafel.*



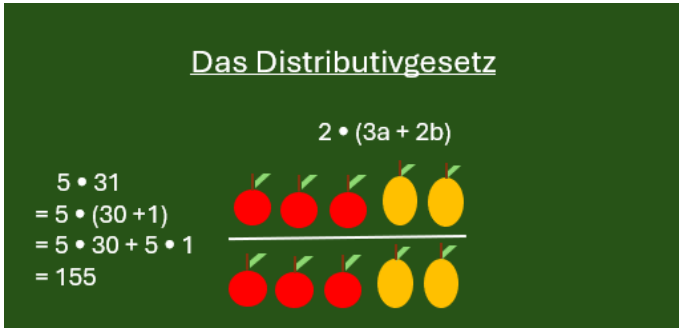

Und den wollen wir vereinfachen.
Was ist denn da das Problem?

*Es melden sich Ana, Theo und Tim.*

Tim?

TIM
Wir wissen halt nicht wie viel 3a+2b ist.

LEHRKRAFT
Genau, das ist das Problem. a und b sind ja Variablen.
Ihr könnt euch das vorstellen wie mit Äpfeln und Birnen. Die darf man ja auch nicht einfach zusammenrechnen.
Machen wir uns mal ein Bild dazu. Weil wir mit zwei multiplizieren, sagt der Term ja, dass wir zwei Mal jeweils drei Äpfel und zwei Birnen haben. Die Variable a könnte ja zum Beispiel für Apfel und die Variable b für Birne stehen.




*Die Lehrkraft zeichnet während des Sprechens zeilenweise erst 3 Äpfel und 2 Birnen, eine horizontale Linie und nochmal 3 Äpfel und 2 Birnen.*

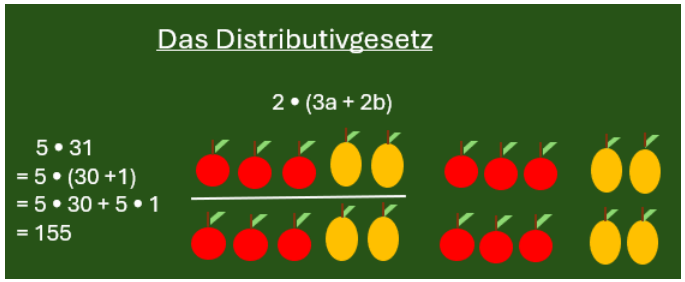
Könnt ihr jetzt sehen, wie viele Äpfel und Birnen wir jetzt insgesamt haben?

*Es melden sich Steffen, Levin, Sophia und Maja.*

Steffen?

STEFFEN
Naja, das sind 6 Äpfel und 4 Birnen.

LEHRKRAFT
Aha. Ich kann das Ganze also auch so zeichnen: 6 Äpfel und 4 Birn en.




*Die Lehrkraft zeichnet während des Sprechens eine zweite Anordnung an die Tafel: Erst alle Äpfel, dann die Birnen.*

Wie können wir den Term also vereinfachen?

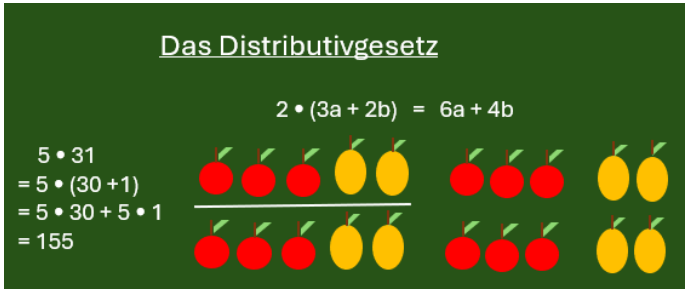
*Es melden sich Steffen, Sara, Levin und Maja.*

Sara?

SARA
6a+4b.

LEHRKRAFT
Sehr gut.

*Die Lehrkraft schreibt an die Tafel und vervollständigt die Gleichung.*




Das heißt, wir können den Term 2•(3a+2b) vereinfachen, indem wir 6a+4b draus machen, weil beide Terme das gleiche Obst beschreiben. Schaut euch mal an, was wir da gemacht haben: 2•3a und 2•2b.
Wir multiplizieren den Faktor vor der Klammer mit den beiden Summanden in der Klammer.

*Die Lehrkraft zeigt in der Gleichung an der Tafel während des Sprechens.*

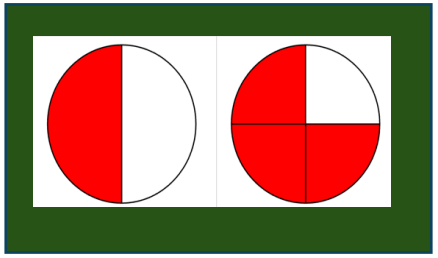
Genau wie ohne Variablen. Das Muster ist das gleiche.
Nur dass wir mit Variablen aufpassen müssen, dass wir Äpfel und Birnen nicht zusammenrechnen dürfen.

**Script for Video 4: Calculation of the area of a circle sector**

Positive example *challenging tasks*

*Beschreibung der Klassensituation:*
*Nachdem sich die Schülerinnen und Schüler in den letzten Unterrichtsstunden mit Kreisen und Berechnungen am Kreis beschäftigt haben, soll in dieser Unterrichtsstunde die Berechnung von Flächeninhalten von Kreisausschnitten behandelt werden. Der folgende Unterrichtsausschnitt findet unmittelbar nach der Begrüßung der Schülerinnen und Schüler statt.*

LEHRKRAFT
Wir haben uns in den letzten Stunden schon ausführlich mit Kreisen und auch Berechnungen am Kreis beschäftigt. Heute geht's drum herauszufinden, wie wir Flächeninhalte von Kreisausschnitten berechnen können. Dazu sollt ihr euch jetzt erstmal überlegen, wie ihr die Flächeninhalte in diesen beiden Beispielen berechnen könnt. Ihr bekommt dazu von mir noch eine Angabe: Beide Kreise haben einen Radius von 10 cm.



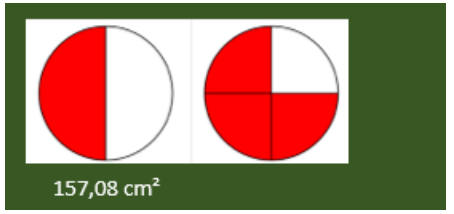
*Die Schülerinnen und Schüler arbeiten in Partnerarbeit zusammen und überlegen, wie sie die Flächeninhalte der Kreisausschnitte berechnen können (ca. 5-10 Minuten).*

LEHRKRAFT
Gut, ich sehe, dass die meisten schon Ideen für die Lösungen aufgeschrieben haben. Dann lasst uns mal eure Überlegungen besprechen. Mia, fang doch mal an.

MIA
Ja also der erste Kreis, das ist ein Halbkreis, der hat einen Flächeninhalt von 157,08 cm².

LEHRKRAFT
Genau, das ist richtig.

*Die Lehrkraft notiert das Ergebnis 157,08 cm² an der Tafel.*

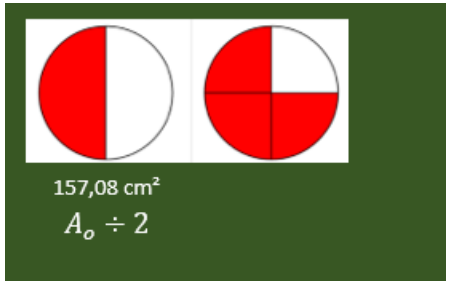



Wie seid ihr denn vorgegangen, um den Flächeninhalt zu berechnen?

*Mia deutet auf den Halbkreis an der Tafel.*

MIA
Ähhm, also bei dem Halbkreis haben wir zuerst die Fläche vom Kreis berechnet. Dann ist das ja aber nur die Hälfte, also haben wir geteilt durch 2 gerechnet.

*Die Lehrkraft nickt lobend und schreibt die Rechnung an die Tafel.*




LEHRKRAFT
Also $A_O$ geteilt durch 2.
Prima.

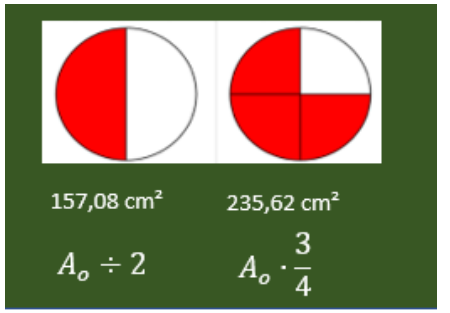
*Die Lehrkraft zeigt auf den Dreiviertel-Kreis.*

Was waren eure Überlegungen bei diesem Kreisausschnitt?

*Es melden sich Mia, Katharina und Paul.*

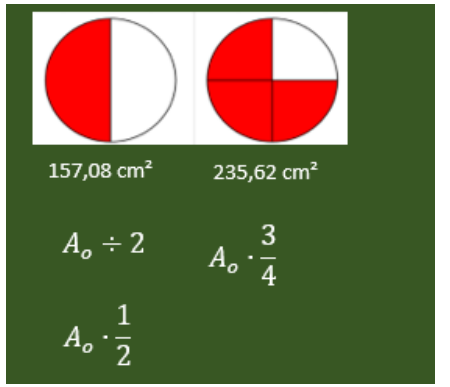
Ja, Katharina.

*Katharina deutet auf den ¾-Kreis.*




KATHARINA
Also man kann mal ¾ rechnen, weil es ja nur drei Viertel vom ganzen Kreis sind. Und der Flächeninhalt ist dann 235,62 cm².

*Lehrkraft ergänzt die Rechnung für den ¾-Kreis an der Tafel.*



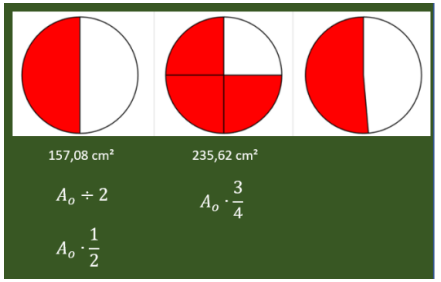

PAUL
Frau Schmidt, können wir das dann links bei dem Halbkreis nicht auch gleich so aufschreiben, also mal ½?

LEHRKRAFT
Ja genau, das schreibe ich gleich noch dazu.

*Lehrkraft ergänzt mal ½ an der Tafel.*

Das habt ihr gut hinbekommen! Jetzt hänge ich hier noch einen weiteren Kreisausschnitt dazu…

*Die Lehrkraft hängt einen weiteren Kreisausschnitt an die Tafel und deutet auf ihn.*




Könnt ihr auch hier den Flächeninhalt bestimmen?

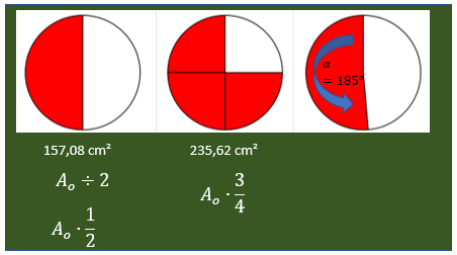
*Die Lehrkraft wartet einen Moment. Es melden sich Julia und Mia.*

Ja, Julia.

JULIA
Hmm, naja, man sieht bei dem Kreisausschnitt eigentlich nur, dass er ein bisschen größer als der Halbkreis ist… Wie man das dann genau berechnet, weiß ich aber auch nicht.

LEHRKRAFT
Mhm. Haben die anderen da Ideen?
*Es melden sich Paul, Malin, Mia und Katharina. Paul wird aufgerufen.*

PAUL
Wir müssten wissen, wie groß der Teil vom ganzen Kreis genau ist.




LEHRKRAFT
Aha…

*Die Lehrkraft zeichnet dann ohne weiteren Kommentar Alpha ein mit 185 Grad. Malin und Mia melden sich nach mehreren Sekunden.*

LEHRKRAFT
Malin.

MALIN
Ahhhjaa, mit Grad geht das. Ein halber Kreis hat ja 180 Grad. Und ähm man sieht ja, dass es ein bisschen mehr ist, also 185 Grad kommt hin. (5 Sekunden Pause)

LEHRKRAFT
Stimmt das? Was sagen die anderen dazu?

*Die Lehrkraft macht eine Pause und wartet einen Moment ab. Es melden sich Mia und Paul.*

MIA
Ähm ja, der ganze Kreis hat ja 360 Grad.

LEHRKRAFT
Hilft uns das, um das Problem von Paul zu lösen? Also wissen wir jetzt, wie groß der Teil vom Ganzen genau ist?

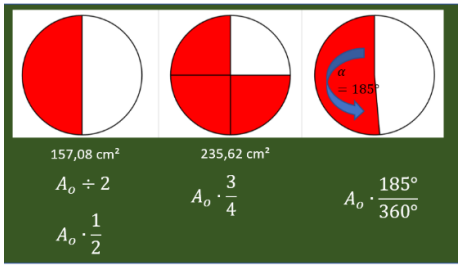
*Die Lehrkraft macht eine kleine Pause. Julia und Hannah melden sich und Hannah wird aufgerufen.*

HANNAH
Sind das jetzt nicht 185 von 360, so wie beim zweiten Beispiel 3 von 4?

LEHRKRAFT
Aha. Kannst du den anderen erklären, wie du das meinst?

*Die Lehrkraft dreht sich halb zur Tafel und schaut Hannah auffordernd an.*

HANNAH
Also wir müssten dann wieder den ganzen Flächeninhalt nehmen und dann mal 185/360 rechnen. Also $A_O$ • 185/360.




LEHRKRAFT
Sind damit alle einverstanden?

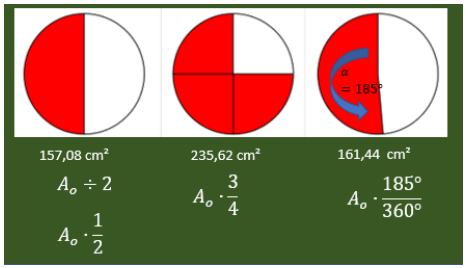
*Mehrere Schülerinnen und Schüler nicken. Die LK schaut in die Runde und wartet kurz ab. Die Lehrkraft ergänzt das Tafelbild.*

Was können wir als Rechnung aufschreiben?

HANNAH
$A_O$ • 185/360.

LEHRKRAFT
Gut. Was kommt denn da raus?

*Ein paar Schülerinnen und Schüler tippen in den Taschenrechner. Dann melden sich Mia, Malin, Julia, Paul, Hannah und Franziska. Die Lehrkraft ruft Franziska auf.*




FRANZISKA
161,44 cm².

*Die Lehrkraft nickt lobend und schreibt das Ergebnis an die Tafel.*

LEHRKRAFT
Und wenn der Mittelpunktswinkel jetzt nicht 185 Grad ist, sondern irgendein anderer Winkel Alpha, z.B. 19 Grad oder 125 Grad? Wie könnten wir dann den Flächeninhalt vom dazugehörigen Kreisausschnitt berechnen?

*Nach mehreren Sekunden melden sich Mia, Malin, Julia und Katharina. Die Lehrkraft wartet ab und ruft dann Malin auf.*

MALIN
Dann ersetzt man eben die 185 durch die andere Zahl.

LEHRKRAFT
Katharina.

KATHARINA
Ja genau, man kann einfach wieder so rechnen wie eben. Also $A_O$ • 19/360, z.B.

LEHRKRAFT
Aha. Und geht das denn immer so?

*Kurze Pause. Mia, Malin, Julia und Katharina melden sich, Julia wird von der Lehrkraft aufgerufen.*

JULIA
Wahrscheinlich schon. (*Schaut die Beispiele an der Tafel an*). Hmm, wobei wir es zuerst ja anders gemacht haben.

LEHRKRAFT
Mia.

MIA
Ich glaube, dass es bei den Kreisausschnitten, die ein ¾-Kreis oder ein Halbkreis sind, es eigentlich auch so geht mit den 360stel. Das wären beim Halbkreis dann ja einfach 180/360.

LEHRKRAFT
Okay. (*Die Lehrkraft macht eine kurze Pause*).
Können wir mit diesen Überlegungen jetzt eine allgemeine Formel herleiten, mit der wir den Flächeninhalt von beliebigen Kreisausschnitten berechnen können? Schaut Euch auch nochmal unsere Beispiele an der Tafel an.

*Nach ein paar Sekunden meldet sich Hannah und wird nach einer kurzen Pause von der Lehrkraft aufgerufen.*

HANNAH
Man berechnet immer erst den Flächeninhalt vom ganzen Kreis.

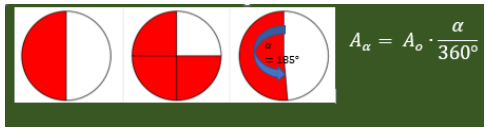
*Katharina meldet sich und wird von der Lehrkraft aufgerufen.*

KATHARINA
Dann hängt es von dem Winkel ab, den der Kreisausschnitt hat.

LEHRKRAFT
Ja, genau. Wir können diesen Winkel einfach Alpha nennen. Wie geht es dann weiter?

*Die Lehrkraft macht eine kurze Pause. Es melden sich Malin, Julia, Franziska und Paul. Die Lehrkraft ruft Malin auf.*

MALIN
Alpha durch 360, also als Bruch geschrieben.α/360, weil wir brauchen ja immer den passenden Teil vom ganzen Kreis.



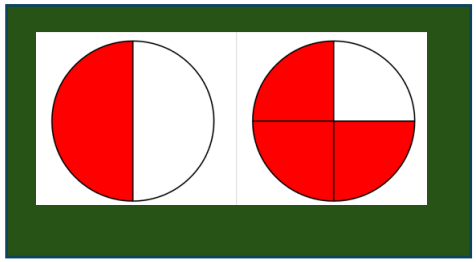

*Die Lehrkraft nickt lobend und schreibt die Formel an die Tafel (siehe Tafelbild) und fasst die Formel abschließend zusammen.*

**Script for Video 5: Calculation of the area of a circle sector**

Negative example *challenging tasks*

*Beschreibung der Klassensituation:*
*Nachdem sich die Schülerinnen und Schüler in den letzten Unterrichtsstunden mit Kreisen und Berechnungen am Kreis beschäftigt haben, soll in dieser Unterrichtsstunde die Berechnung von Flächeninhalten von Kreisausschnitten behandelt werden. Der folgende Unterrichtsausschnitt findet unmittelbar nach der Begrüßung der Schülerinnen und Schüler statt.*

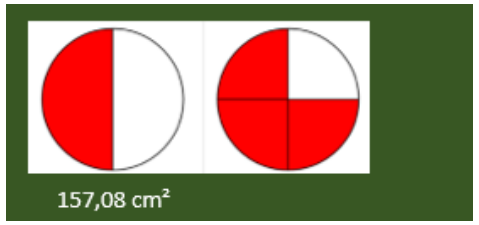
LEHRKRAFT
Wir haben uns in den letzten Stunden schon ausführlich mit Kreisen und auch Berechnungen am Kreis beschäftigt. Heute geht's drum herauszufinden, wie wir Flächeninhalte von Kreisausschnitten berechnen können. Dazu sollt ihr euch jetzt erstmal überlegen, wie ihr die Flächeninhalte in diesen beiden Beispielen berechnen könnt. Ihr bekommt dazu von mir noch eine Angabe: Beide Kreise haben einen Radius von 10 cm.



*Die Schülerinnen und Schüler arbeiten in Partnerarbeit zusammen und überlegen, wie sie die Flächeninhalte der Kreisausschnitte berechnen können (ca. 5-10 Minuten).*

LEHRKRAFT
Gut, ich sehe, dass die meisten schon Ideen für die Lösungen aufgeschrieben haben. Dann lasst uns mal eure Überlegungen besprechen. Mia, fang doch mal an.

MIA
Ja also der erste Kreis, das ist ein Halbkreis, der hat einen Flächeninhalt von 157,08 cm².




LEHRKRAFT
Genau, das ist richtig.

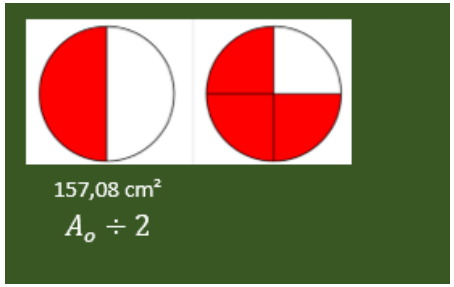
*Die Lehrkraft notiert das Ergebnis 157,08 cm² an der Tafel.*

Wie seid ihr denn vorgegangen, um den Flächeninhalt zu berechnen?

*Mia deutet auf den Halbkreis an der Tafel.*

MIA
Ähhm, also bei dem Halbkreis haben wir zuerst die Fläche vom Kreis berechnet. Dann ist das ja aber nur die Hälfte, also haben wir geteilt durch 2 gerechnet.

*Die Lehrkraft nickt lobend und schreibt die Rechnung an die Tafel.*




LEHRKRAFT
Also $A_O$ geteilt durch 2.
Prima.

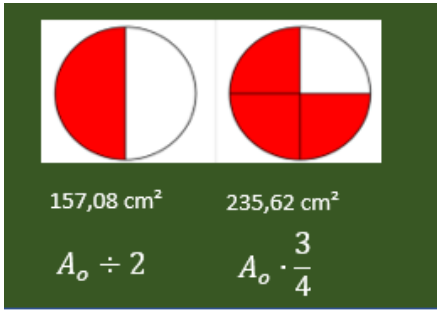
*Die Lehrkraft zeigt auf den Dreiviertel-Kreis.*

Was waren eure Überlegungen bei diesem Kreisausschnitt?

*Es melden sich Mia, Katharina und Paul.*

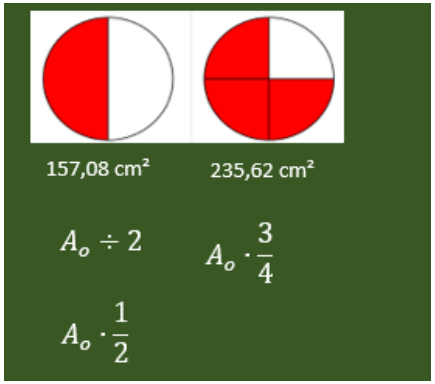
Ja, Katharina.

*Katharina deutet auf den ¾-Kreis.*




KATHARINA
Also man kann mal ¾ rechnen, weil es ja nur drei Viertel vom ganzen Kreis sind. Und der Flächeninhalt ist dann 235,62 cm².

*Lehrkraft ergänzt die Rechnung für den ¾-Kreis an der Tafel.*



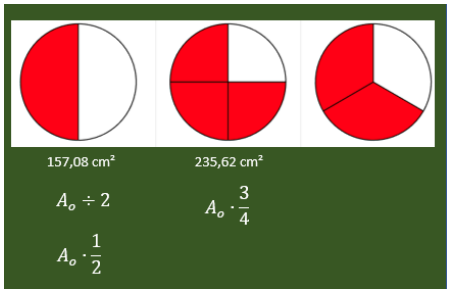

PAUL
Frau Schmidt, können wir das dann links bei dem Halbkreis nicht auch gleich so aufschreiben, also mal ½?

LEHRKRAFT
Ja genau, das schreibe ich gleich noch dazu.

*Lehrkraft ergänzt mal ½ an der Tafel.*

Das habt ihr gut hinbekommen! Jetzt hänge ich hier noch einen weiteren Kreisausschnitt dazu…

*Die Lehrkraft hängt einen weiteren Kreisausschnitt an die Tafel und deutet auf ihn.*




Könnt ihr auch hier den Flächeninhalt bestimmen?

*Die Lehrkraft wartet einen Moment. Es melden sich Julia und Mia.*

Ja, Julia.

JULIA
Hmm, ich glaube den kann man ähnlich berechnen. Das sind ja 2/3. Also einfach den Flächeninhalt vom ganzen Kreis mal 2/3.

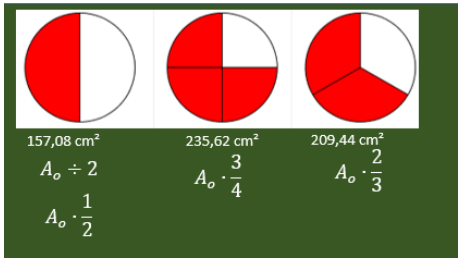
LEHRKRAFT
Mhm. Das schreiben wir gleich mal dazu.

*Die Lehrkraft schreibt den Term unter den Kreisausschnitt an der Tafel.*

Was kommt dann da raus?

*Mehrere Schülerinnen und Schüler tippen in den Taschenrechner und es melden sich Paul, Malin und Franziska.*

Ja, Paul.




PAUL
Das sind dann 209,44 cm².

LEHRKRAFT
Genau.

*Die Lehrkraft schreibt das Ergebnis an die Tafel.*

Dann sammeln wir jetzt mal noch weitere Kreisausschnitte, damit wir herausfinden, wie wir den Flächeninhalt allgemein berechnen können. Habt ihr Ideen für Kreisausschnitte?

*Mia, Malin, Julia und Paul melden sich nach mehreren Sekunden.*

LEHRKRAFT
Malin.

MALIN
Zum Beispiel einen Fünftelkreis, da würden wir dann geteilt durch 5 rechnen.

*Tippt nebenher in den Taschenrechner ein.*

Da kommt dann 62,83 cm² raus.

LEHRKRAFT
Gut. Können wir die Rechnung auch so aufschreiben, wie bei den anderen Beispielen?
Mia.

MIA
Ja, $A_O$ • 1/5 wäre das dann.

LEHRKRAFT
Mhm. Habt ihr noch andere, vielleicht auch schwierigere Beispiele?

*Die Lehrkraft macht eine kleine Pause. Hannah und Mia melden sich und Hannah wird aufgerufen.*

HANNAH
Ich hab noch 3/7. Das ist schwieriger zu zeichnen.

LEHRKRAFT
Aha. Und ist der Flächeninhalt dann auch schwieriger zu berechnen?

*Es melden sich Mia und Julia und Julia wird aufgerufen.*

JULIA
Ja, also man kann halt nicht einfach durch eine Zahl teilen.

LEHRKRAFT
Was sagen die anderen?

*Es meldet sich Mia und wird aufgerufen.*

MIA
Man kann ja mal 3/7 rechnen. So wie vorhin bei mal ¾.

LEHRKRAFT
Sind damit alle einverstanden?

*Mehrere Schülerinnen und Schüler nicken. Die Lehrkraft schaut in die Runde und wartet kurz ab.*

Habt ihr noch andere Beispiele?

*Es melden sich Franziska, Malin, Julia und Paul. Die Lehrkraft ruft Franziska auf.*

FRANZISKA
Wenn man einen 1/10 Kreis hätte oder einen 9/10 Kreis, dann müsste es auch genauso gehen. Also $A_O$ mal 1/10 oder halt mal 9/10.

*Die Lehrkraft nickt lobend.*

LEHRKRAFT
So, jetzt haben wir ja einige Beispiele gesammelt. Wenn wir jetzt aber irgendeinen beliebigen Kreisausschnitt haben, mit irgendeinem Mittelpunktswinkel Alpha, wie könnten wir dann den Flächeninhalt berechnen?

*Nach mehreren Sekunden melden sich Mia, Malin, Franziska und Katharina. Die Lehrkraft wartet ab und ruft Malin dann auf.*

MALIN
Dann ersetzt man eben zum Beispiel die 2/3 (*zeigt auf das Beispiel an der Tafel*) durch den neuen Bruch.

LEHRKRAFT
Franziska.

FRANZISKA
Man kann doch immer zum Beispiel durch 3 rechnen bei einem Drittelkreis und dadurch den Flächeninhalt von jedem beliebigen Kreisausschnitt berechnen.

Die *Lehrkraft wartet ab und es melden sich Mia und Malin. Mia wird von der Lehrkraft aufgerufen.*

MIA
Also ich glaube man kann einfach immer so rechnen: den Flächeninhalt vom ganzen Kreis mal dem Teil vom ganzen Kreis, den wir berechnen wollen.

LEHRKRAFT
Okay… (*Die Lehrkraft macht eine kurze Pause*). Können wir mit diesen Überlegungen jetzt eine allgemeine Formel herleiten, mit der wir den Flächeninhalt von beliebigen Kreisausschnitten berechnen können? Schaut Euch auch nochmal unsere Beispiele an der Tafel an.

*Nach ein paar Sekunden meldet sich Hannah und wird nach einer kurzen Pause von der Lehrkraft aufgerufen*.

HANNAH
Man berechnet immer erst den Flächeninhalt vom ganzen Kreis.

*Katharina meldet sich und wird von der Lehrkraft aufgerufen.*

KATHARINA
Und dann muss man halt zum Beispiel mal ½ rechnen bei einem Halbkreis.

LEHRKRAFT
Genau. Und beim Halbkreis ist ja der Mittelpunktswinkel 180 Grad. Und 1/2 kann man ja auch schreiben als 180/360. Denn der ganze Kreis hat 360 Grad, das hatten wir ja schon besprochen. Und wenn wir jetzt irgendeinen beliebigen Mittelpunktswinkel alpha haben?

*Die Lehrkraft macht eine kurze Pause. Es melden sich Malin, Julia, Franziska und Paul. Malin wird nach ein paar Sekunden aufgerufen.*

MALIN
Dann müssen wir die 180 Grad durch Alpha ersetzen?

*Die Lehrkraft nickt lobend und schreibt die Formel an die Tafel (siehe Tafelbild) und fasst die Formel abschließend zusammen.*