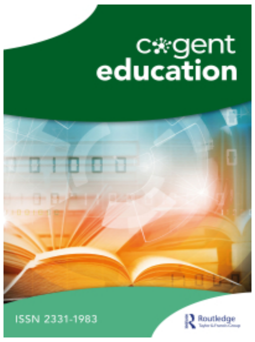


Taylor & Francis
Taylor & Francis Group

Cogent Education

ISSN: 2331-186X (Online) Journal homepage: www.tandfonline.com/journals/oaed20

Developing science teachers' enacted pedagogical content knowledge through integration of student feedback into the Refined Consensus Model of pedagogical content knowledge

Daniel Hartmuth, Rifat Mehreen Amin, Daniel Schlichting, Svenja Schött, Jonathan Haudenschild, Helmut Küchenhoff & Birgit J. Neuhaus

To cite this article: Daniel Hartmuth, Rifat Mehreen Amin, Daniel Schlichting, Svenja Schött, Jonathan Haudenschild, Helmut Küchenhoff & Birgit J. Neuhaus (2025) Developing science teachers' enacted pedagogical content knowledge through integration of student feedback into the Refined Consensus Model of pedagogical content knowledge, Cogent Education, 12:1, 2470563, DOI: 10.1080/2331186X.2025.2470563

To link to this article: https://doi.org/10.1080/2331186X.2025.2470563

Published online: 01 Mar 2025.

COGENT EDUCATION
2025, VOL. 12, NO. 1, 2470563
https://doi.org/10.1080/2331186X.2025.2470563

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE
OPEN ACCESS Check for updates

Developing science teachers' enacted pedagogical content knowledge through integration of student feedback into the Refined Consensus Model of pedagogical content knowledge

Daniel Hartmuth[a], Rifat Mehreen Amin[b], Daniel Schlichting[c], Svenja Schött[b], Jonathan Haudenschild[b], Helmut Küchenhoff[c] and Birgit J. Neuhaus[a]

[a]Faculty for Biology, LMU Munich, Munich, Germany; [b]Faculty for Informatics, LMU Munich, Munich, Germany; [c]Department of Statistics, LMU Munich, Munich, Germany

ABSTRACT
Developing Pedagogical Content Knowledge (PCK) is a facet of teachers' professional competence. The Refined Consensus Model of PCK (RCM) emphasises a cyclical process structure to develop an individual's enacted PCK (ePCK) and the plan-teach-reflect cycle must be systematically adjusted by means of feedback loops. Simple graphic visualisation of student feedback over time provide the information for teachers to improve ePCK. The research gap lies in providing teachers with an actionable cyclic feedback process and evaluating its effectiveness. In an intervention study with 28 in-service science teachers and 137 lessons at secondary schools in Bavaria, Germany, we analysed the effect of student feedback for developing ePCK. We compared two types of visual feedback for how they influenced the teacher's ability in the classroom, indirectly measured as a change in the students' assessment of lesson quality. Lesson ratings improved slightly but not significantly when teachers used the feedback tool. The high intra-class correlation of consecutive lesson assessments supported the use of rolling average curves in the graphics. We conclude that science teachers can foster ePCK with the help of student feedback, which empowers teachers in practice and serves as a valuable tool in research to investigate the development of ePCK in greater detail.

IMPACT STATEMENT
This study shows that science teachers can foster ePCK with the help of student feedback, which not only empowers teachers in practice but also serves as a valuable tool in research to investigate the development of ePCK in greater detail.

ARTICLE HISTORY
Received 18 September 2024
Revised 23 January 2025
Accepted 18 February 2025

KEYWORDS
RCM; ePCK; plan-teach-reflect cycle; student feedback; instructional quality; biology education; teacher education; higher education

SUBJECTS
Teacher Training; Teachers & Teacher Education; Teacher Education & Training

Introduction

Developing teaching strategies and the process of teaching are dynamic processes (Hattie, 2009), with a myriad of interrelated variables and interdependent non-linear effects on teaching outcomes (Gess-Newsome & Lederman, 1999; Neuhaus, 2021).

Because many variables contributing to teaching skills are entangled (Alonzo et al., 2019; Carlson et al., 2019; Mientus et al., 2022), and difficult to measure (Park, 2019) teachers have found difficulty implementing scientifically acknowledged successful teaching practices (Behling et al., 2022; Shavelson, 2020). In order to develop practical teaching skills, a cyclical feedback process is needed that can provide useful information for an individual teacher's plan-teach-reflect cycle (Carlson et al., 2019). In recent years, various methods have been employed to establish this cycle by digital technology, offering educators in classroom settings feedback on their teaching methods (Suresh et al., 2021). One practical method for acquiring an accurate assessment of teaching effectiveness is through student feedback (Wisniewski et al., 2020), and which is recognised as a significant determinant of teaching excellence (Hattie & Timperley, 2007). Focusing on student feedback for educators through digital

CONTACT Daniel Hartmuth Daniel.hartmuth@biologie.uni-muenchen.de Faculty for Biology, LMU Munich, Munich, Germany

rating systems, which allow feedback to be given with a single click and delivered as an aggregate score, enables swift and highly practical responses (Bijlsma et al., 2022), preventing the loss of valuable insights otherwise obtained through extensive feedback questionnaires (Wisniewski & Zierer, 2020). With student feedback in particular, a compromise must be found between the time spent collecting feedback, the complexity and number of the questions and the degree of information provided by the feedback (Wisniewski & Zierer, 2021). In order to be able to use student feedback to ensure comparability between teachers or between scientific studies, the tendency has often been to collect student feedback with low frequency and with detailed feedback questions (Hattie, 2009). This trend has created a clear research gap that focuses on the cyclical application of student feedback. By opting to enhance individual teaching methodologies gradually, we chose to explore a method involving a unified student feedback score obtained through a single student evaluation (Rollett et al., 2021). By employing rapid and generalised student feedback, we applied such noise reduction methods as aggregation, moving-average curves and relative data visualisation to highlight the evolving nature of practical teaching skills to teachers (Kahneman et al., 2021).

The goal of this study was to implement student feedback for teachers into the plan-teach-reflect-cycle and thus enable successful development of teaching skills. We developed a digital tool that enables the collection of student lesson ratings in a dynamic classroom environment, and the presentation of the results over time as a graphical visualisation of this feedback. Two visual representations illustrated the temporal progression of mean ratings for a particular class. Similar to performance and development-orientated visualisations in finance, one displays a moving-average curve, while the other depicts a coordinate plane with the slope of relative performance on the x-axis and relative performance on the y-axis. This approach enables teachers to monitor the ratings of current versus prior lessons and to adapt their teaching methods to the student ratings, thus developing their teaching skills.

To provide a comprehensive understanding of the theoretical foundations, the following section systematically explores the key components: First, the evolution of theories on Pedagogical Content Knowledge (PCK) is examined. Second, the historical shift towards a dynamic interpretation of PCK is highlighted, culminating in the formulation of the Refined Consensus Model (RCM). Third, the conditions necessary for the successful development of teachers' ePCK and the methods for assessing it are discussed. Finally, the role of student feedback in fostering the development of both ePCK and pPCK is analysed, and the research objectives and guiding questions of this study are presented. Together, these interconnected elements form the foundation for exploring how student feedback-driven strategies can enhance teaching effectiveness in dynamic classroom settings.

Theory

The evolution of PCK

Since the concept of pedagogical content knowledge (PCK) was coined 35 years ago, understanding how subject-specific, experience-based knowledge forms and transforms other realms of knowledge (pedagogical, content) has increased significantly (Shulman, 1986). A major contribution to the success of the PCK concept is its emphasis on subject-specific student needs and flexible teaching strategies. For instance, addressing preconceptions significantly influences student outcomes in science classes (Driver et al., 1994).

The scientific models constructed around or integrating PCK have undergone a fascinating evolution (see Figure 1) that integrates empirical findings and adapts to the needs of both practitioners and science education researchers (Carlson et al., 2019).

PCK encompasses the interplay of content, action and teacher rationale, with a particular emphasis on enhancing topic-specific student outcomes. Beyond this, it also considers factors such as fostering students' motivation and enjoyment of learning, reflecting its broader role in shaping effective teaching practices (Berry et al., 2015). Discussions often revolve around PCK's distinctiveness compared to pedagogical and content knowledge (Magnusson et al., 1999). PCK's transformation of various knowledge types through action underscores its dynamic nature and unique professional role for teachers (Shulman, 1987). Despite unclear boundaries, PCK's appeal lies in defining competencies required for expert science teachers and influencing related research (Marks, 1990; Shulman, 1987).

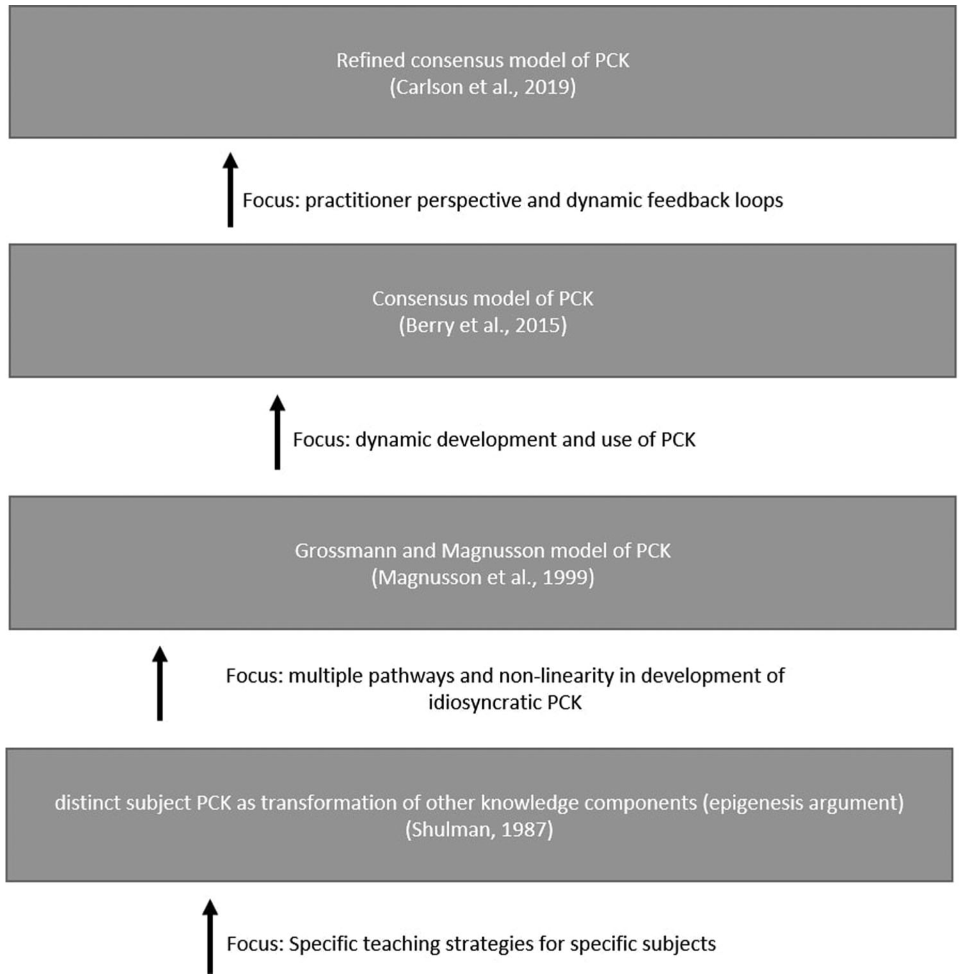


Figure 1. Evolution of the Refined Consensus Model. Initial foci of development are included in the next evolutionary step of the model. In all stages adaptions are made by the scientific community to adjust for the needs of a changing practitioner environment.

Models like the Grossman and Magnusson frameworks guide teachers in embedding scientifically meaningful learning experiences, illustrating PCK's specificity and non-linearity (Grossman, 1990; Magnusson et al., 1999).

From these models, the consensus model of PCK emerged, emphasising its dynamic nature during teaching and its focus on explicit scientific teaching skills across subjects, topics and concepts (Berry et al., 2015; Gess-Newsome et al., 2019). In its latest consensus version, the model emphasises the practitioner's perspective of PCK and draws further attention to the many feedback loops involved in the development of PCK in practice. The illustration of these feedback loops shows the pedagogical reasoning cycles of instruction, also called the plan-teach-reflect-cycle (Carlson et al., 2019). The different layers of the model are interconnected through filters for and enhancers of knowledge exchange, emphasising the inherent dynamic character and non-linearity of the model. These hypothesised filters and enhancers should be considered as catalysts for knowledge exchange, rather than causal links or correlated factors expected for a linear model (Carlson et al., 2019). Although the RCM focuses on distinct realms within PCK, it acknowledges Shulmans first definition of PCK that the ' (…) broader professional knowledge bases are foundational to teacher PCK' (Carlson et al., 2019).

A dynamic interpretation of the Refined Consensus Model

The RCM situates science teaching practice – via the pedagogical reasoning cycle – at its core. The dynamic aspect of ePCK is emphasised in the whole model; in the paper presenting the model, the word 'dynamic' appears 41 times and within the section on ePCK with 7 mentions alone (Carlson et al., 2019). Three distinct components interact dynamically with situational teaching behaviours: enacted PCK (ePCK), personal PCK (pPCK) and collective PCK (cPCK). Enacted PCK represents situation-specific teacher actions, emphasising its context-dependent and dynamic nature (Carlson et al., 2019). This concept is illustrated by the 'cone' analogy, where ePCK slices represent individual teaching episodes (Figure 2). A complete representation of ePCK, as depicted in Figure 3, shows these slices forming a cohesive structure, signifying the teacher's cumulative teaching potential. This approach is particularly necessary because it captures the dynamic and situation-specific nature of teaching, addressing gaps in static models and providing a more nuanced understanding of teacher actions within diverse contexts. Moreover, it advances the field by proposing a framework that systematically integrates individual teaching episodes into a coherent model of professional growth (Carlson et al., 2019).

This kit of actionable individual skill and knowledge tools, called ePCK, is embedded in the second component of PCK, personal PCK (pPCK).

Personal PCK evolves as teachers articulate aspects of their ePCK, transforming situational insights into conscious knowledge. Conversely, pPCK employed in action becomes ePCK. The dynamic interchange between ePCK and pPCK is illustrated through the 'spinning top' analogy, where pPCK serves as the handle driving the rotation of ePCK, and the spinning top itself represents the interplay between knowledge forms (Figure 4). Moreover, collective PCK, derived from the broader professional community, serves as another source for developing personal PCK (Irmer et al., 2023).

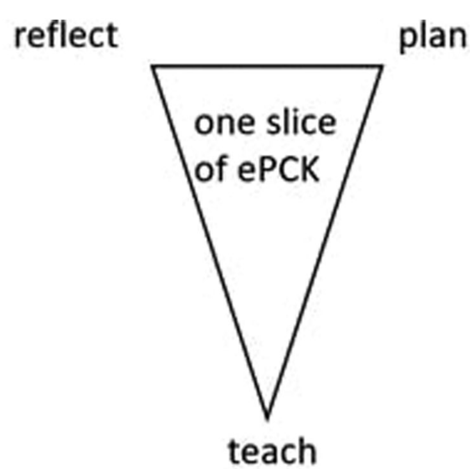


Figure 2. One slice of ePCK. The teaching cycle with its three components illustrated as a triangular slice. The illustration of the slice focuses on a single aspect of ePCK as, for example, dealing with a student error or cognitive activation in a specific situation.

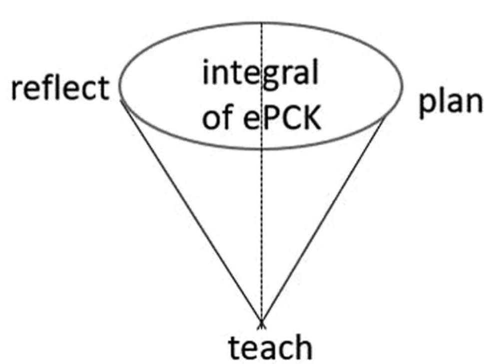


Figure 3. Integral parts of ePCK. The teaching cycle with its three components illustrated as an integral part of ePCK. The situation-specific slices of ePCK taken together form the integral ePCK of an individual teacher, illustrated as a cone.

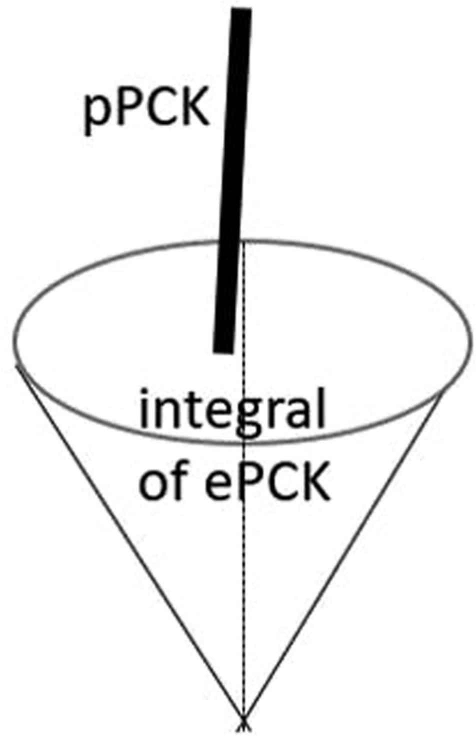


Figure 4. pPCK arising from ePCK. Resembling a spinning top, pPCK emerges as the rotational axis from ePCK, and pPCK can be transformed into the movement (ePCK) of the top.

Not all of the actionable knowledge and skills displayed by a teacher as ePCK in a myriad of teaching situations can be articulated and recognised by the teacher or to any other observer (Behling et al., 2022), but only by means of communication and interaction with the professional community via verbal reasoning (McGuinness & Wittgenstein, 2012). The pedagogical content knowledge of the scientific community as a whole, called collective pedagogical content knowledge (cPCK), forms a second source of information (in addition to ePCK) for the development of pPCK. Just as the relationship between ePCK and pPCK is symmetrical, so is that between pPCK and cPCK. One source of development for cCPK of the whole scientific community is the expressible, articulable body of knowledge of all teachers (Seidel & Shavelson, 2007).

This framework also emphasises the transformative potential of collective insights, ensuring that individual teacher development aligns with broader professional standards and advances collaborative practices in science education.

In terms of the spinning top analogy cPCK can be considered the hand on the top's axis, which emerges from the contribution of many teacher practitioners (Figure 5).

Whereas cPCK is a shared property of many subjects, especially in their scientific community, pPCK and ePCK are particular to each teacher. Moreover, general cPCK cannot be transferred unchanged to an individual pPCK, as the translation and integration depends on the learning context, which comprises all circumstances and factors, such as education policy and curriculum, political influences and many individual attributes (Carlson et al., 2019). The broader learning environment acts as a catalyst, mediating these components' interactions. The learning context can be seen as an environment in which each top is spinning (Figure 6), and just as the top's behaviour is crucially dependent on the surface it spins on, each teacher's individual knowledge and set of skills is mediated by the learning context. It is important to emphasise that the influence of the learning environment on the whole body of PCK does not fit a causal model. The learning environment functions rather like a catalyst, or in the words of the RCM as 'filter and amplification on the feedback loops of the PCK' (Carlson et al., 2019). Catalysts and filters perform differently and dynamically for every teacher and ultimately shape the actual teaching performance in the classroom.

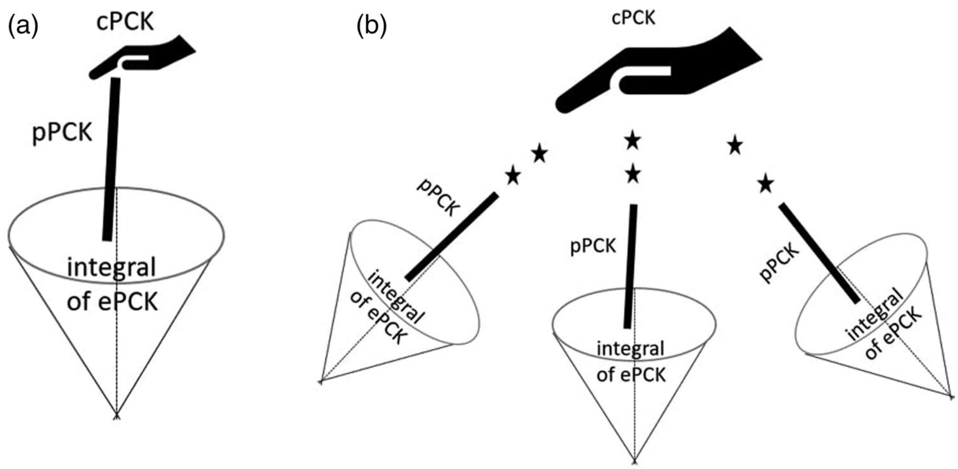


Figure 5. (a) Interrelation of cPCK and personal components of PCK: CPCK seen as a hand driving pPCK. cPCK can be considered a way of gauging a teacher's pPCK and ultimately teaching behaviour in many situations as ePCK. (b) Interrelation of cPCK and personal components of PCK: CPCK fed by pPCK. The contribution of many teachers' pPCK to the body of knowledge of a scientific didactic community in form of cPCK is illustrated by one hand controlling many spinning tops of pPCK and ePCK.

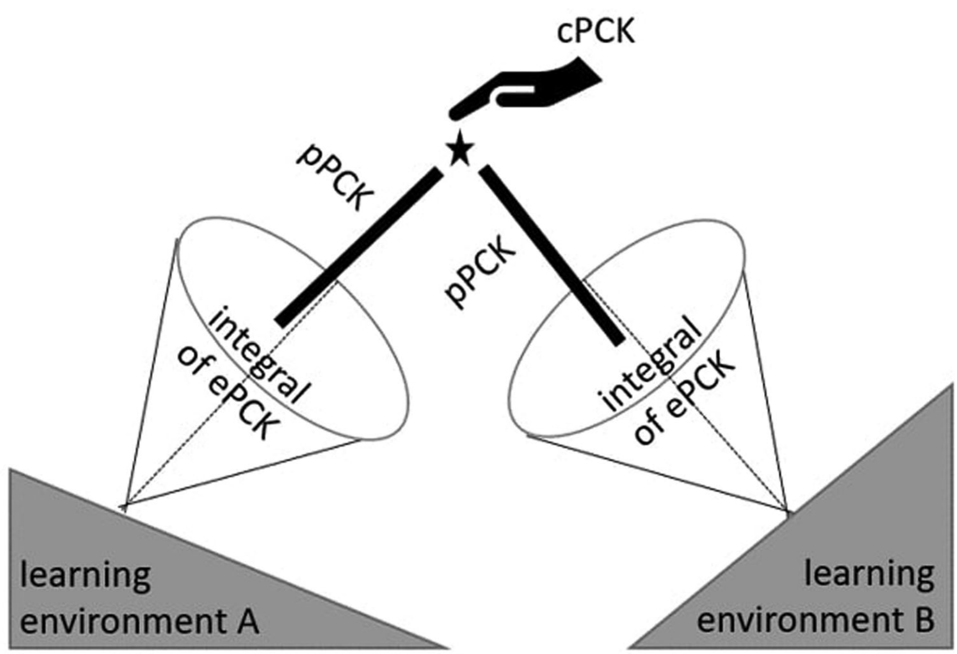


Figure 6. Influence of learning environment on PCK. Under different learning environments A and B, teachers display different sets of ePCK, pPCK which interact differently with cPCK.

Conditions for the successful development of teacher ePCK

The RCM underscores the interdependence of ePCK and pPCK, highlighting their emergence from broader knowledge bases. Research emphasises the importance of integrating fragmented knowledge into deep structures via experience and training (Carlson et al., 2019; Gess-Newsome & Lederman, 1999; Shulman, 1986). From this the question arises of how to foster individual, situation-dependent ePCK and pPCK as concepts that are more than, yet different from, the sum of their constituents (Morine-Dershimer & Kent, 1999).

The current state-of-the-art education curricula for teachers in teacher programs already achieve the development of measurable pPCK and ePCK (Behling et al., 2022; Irmer et al., 2023; Kramer et al., 2020). It is rather the search for, and identification of, factors influencing the successful implementation of actionable pPCK into ePCK across different teaching situations, or sustainable ePCK, that poses difficulties (Behling et al., 2022). This search has found two factors of importance; the learning context or environment and the development of ePCK and pPCK through action and experience (Morine-Dershimer & Kent, 1999; Sabers et al., 1991). Many studies highlight the importance of integration of initially superficial and fragmented pieces of knowledge and skills into internalised deep knowledge structures through training (Lederman & Gess-Newsome, 1999; Tobin & McRobbie, 1999). For the development of ePCK – teaching experience, or at least critical reflection on teaching practices is needed (Davidowitz & Potgieter, 2016); in the absence of practical training of contextualised teaching situations, individual PCK does not improve (Drits-Esser et al., 2017; Förtsch et al., 2016). To a large extent the question of how to improve ePCK in practice systematically remains unclear (Mientus et al., 2022; Sorge et al., 2019). This approach addresses these challenges by advocating for iterative feedback cycles and experiential learning, ensuring that teacher growth is both evidence-based and adaptable to evolving classroom dynamics. Furthermore, it advances current understanding by linking the theoretical constructs of ePCK with practical, measurable outcomes in teaching performance. This theory-to-practice link can further be investigated by drawing on the RCM as a guide. This model highlights two pathways for the development of a teacher's ePCK. First, development by undergoing the teaching cycle in a repetitive feedback loop and second by the interaction of a teacher with the scientific community, accessing the broader cPCK as a source of personal development of ePCK and pPCK (Mientus et al., 2022). These findings and their focus on development of individual PCK through action and experience are consistent with current constructivist learning theories, notably the concept of experiential learning (Kolb & Kolb, 2005; Morris, 2020). Each teacher has to develop ePCK and pPCK in action and dependent on the learning context individually, as illustrated in Figure 7.

Adapting to different and changing learning environments, and dependent on their current status of pPCK and ePCK to draw upon in a given situation, a teacher develops a repertoire of skills and knowledge that improves performance in any environment, upgrading his ePCK and pPCK (Tobin & McRobbie, 1999). can only do so within the constraints of his own learning environment and set of pPCK and ePCK (Shavelson, 2020). If there is a misfit between these constraints and a teaching strategy, teachers consistently fail to incorporate the intended interventions (Behling et al., 2022; Felser, 2015; Shavelson, 2020). As illustrated in Figure 8, a teacher performs as well as he can within all constraints to action (punctuated equilibrium A), but for them to integrate different, new or successful teaching strategies (punctuated equilibrium B) it is only rational and feasible for a compatible individual when there is an incentive (Gould, 2007).

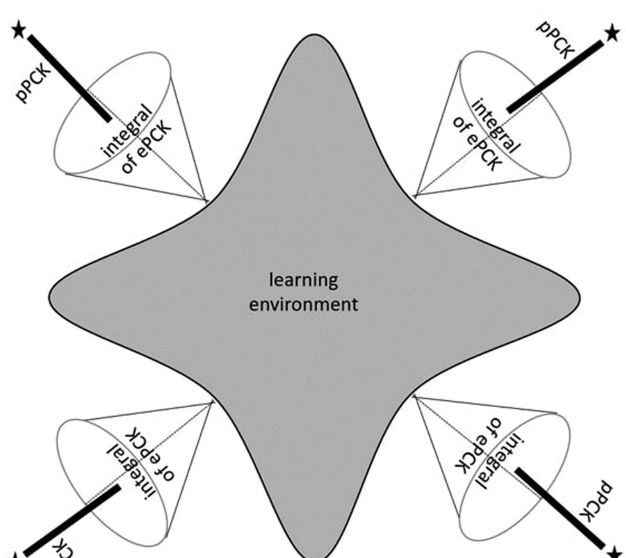


Figure 7. Idiosyncrasy of PCK. The learning environment is different for each teacher, catalysing a different ePCK and pPCK for each individual.

Possible adaptations must fit into this dynamic set of variables, or be discarded at the individual level (Taleb, 2018). This reasoning requires convincing individual teachers of the need for adaptations to their teaching actions through the manifestation of success. These insights contribute to advancing the field by demonstrating how targeted interventions can transform fragmented knowledge into cohesive teaching strategies, driving both theoretical and practical advancements in science education. The iterative feedback loop central to this process not only strengthens teacher competencies but also fosters a deeper integration of research insights into practical application. However, to demonstrate the success of their adapted teaching strategies to teachers, their ePCK must be dynamically assessed.

Assessment of teachers' ePCK

Whether a teacher adapts his ePCK and pPCK in response to external factors (cPCK or learning environment) or experience, and whether this is done consciously (pPCK) or in deed (ePCK), he needs reliable feedback from the environment about the success of the ePCK employed. If this central requirement is not fulfilled, a teacher has no chance of developing ePCK more than fortuitously (Skinner, 1959). There is a wide range of methods and scientifically grounded benchmarks to supply this kind of actionable information for both teachers and science education researchers (Baumert et al., 2010; Blömeke et al., 2014; Mientus et al., 2022).

The most pre-eminent benchmark and proxy of teachers' ePCK success is via student outcome. It has been shown that ePCK has a direct influence (Alonzo et al., 2012) on student outcomes (M. M. Keller et al., 2017) and student competence (Mahler et al., 2017).

Another proxy is via observing teaching in action. This can be done by experts, colleagues, deep-learning engines or students in class. During the practical stage of teacher training it is often expert colleagues who assess ePCK on the basis of empirically founded criteria (Kramer et al., 2020; Wang et al., 1993). During the more theoretical stages of such training individual PCK is often targeted via pen-and-pencil articulation of pPCK in connection with video-based teaching situations (Irmer et al., 2023; Seidel & Stürmer, 2014; Shavelson, 2012; Wang et al., 1993), but there are several problems associated with trying to measure pPCK. First, there is not yet a consensus on what the construct encompasses or where the boundaries of pPCK in contrast to other constructs lie, because empirical research on assessing pPCK and ePCK is preliminary (Abell, 2008; Kirschner et al., 2015). Second, instruments designed to measure pPCK empirically, may measure related constructs such as beliefs about instruction instead (Smith et al., 2015).

Another option for teachers' ePCK assessment is to ask students in class (Kämpfe, 2009) observe teachers in action (Hattie & Clarke, 2019), since recourse to expert observers or advanced technological methods often requires unfeasible resources (Bijlsma et al., 2022; Chan & Yung, 2018; Hense, 2013). The

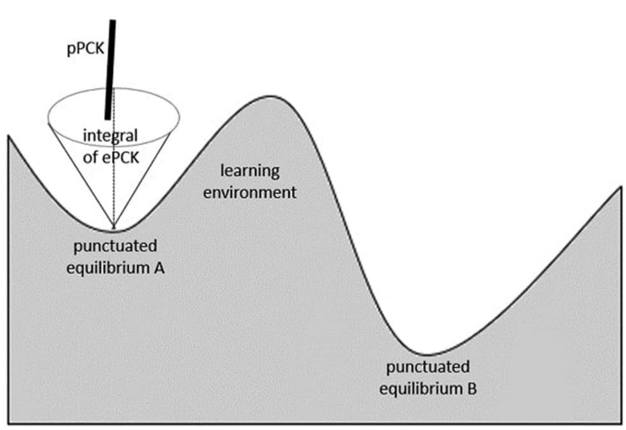


Figure 8. Metastability of individual PCK. A top's spinning behaviour (teaching) depends on inner (ePCK and pPCK) and outer (learning environment) constraints. In actual situations, tops and teachers adapt to achieve a state of punctuated equilibrium.

RCM emphasises the importance of in-practice development ePCK and pPCK at an individual level, with the teaching cycle at its centre (Blömeke & Kaiser, 2014). A feedback loop of reflection, planning and teaching requires recurring observational information about a teacher's progress to provide the informational needs of the teaching cycle (Carlson et al., 2019; Chan & Hume, 2019). This alignment between assessment and actionable insights not only advances the theoretical framework of PCK but also provides practical tools for enhancing teacher effectiveness. However, this raises the pivotal question of identifying the most suitable provider of such feedback.

Student feedback in the context of developing ePCK and pPCK

For measuring teaching quality and development objectively subject-specific concepts are often used because a focus on these has provided robust results (Bijlsma et al., 2021; Neuhaus, 2021). The assessment of teaching quality requires clear criteria (Wisniewski & Zierer, 2021) and the effects of student feedback on teacher improvement are large, with effect sizes in meta-analyses of around $d = 0.7$ (Hattie & Timperley, 2007) makes student feedback an important tool for improving teachers' lesson quality and professional competence (DeWitt, 2018; Hattie & Timperley, 2007). The strong effect of student feedback is well established (Gaertner, 2014; Helmke & Lenske, 2013; Hense, 2013) and a key requirement for goal-orientated reflection of teaching practices (Wisniewski et al., 2020), which can be improved by focusing on process-orientated feedback (Bastian et al., 2016; Brühwiler et al., 2017). Thus teaching development and iterative adaptations in teaching behaviour interact cyclically (Bastian et al., 2016).

Only by viewing contributing factors together can they be organised effectively to improve the quality of teaching (Praetorius et al., 2012); indeed the RCM outlines this procedure with a central cyclical structure and emphasises teaching skill evolution as a process resembling a feedback loop, as viewed through research in the natural sciences (Behling et al., 2022). By sequentially following this plan-teach-reflect-cycle, a teacher develops additional ePCK and pPCK, putting reflected pedagogical content knowledge into action (Irmer et al., 2023). The pedagogic literature on effective student feedback, and the RCM with its subject-specific background, converge on a cyclical structure, in which reflection is intertwined with lesson planning and the act of teaching. In various studies, however, teachers fail to integrate information on faults or improvements to their teaching practices (Shavelson, 2020). it is neither automatic nor easy to develop ePCK in this way because a trigger is necessary (Behling et al., 2022). Evidence suggests that feedback serves as a catalyst for reflection within the classroom setting, exerting an appreciable influence on lesson preparation and teaching methodology and thus ePCK (Hattie & Clarke, 2019). Integrating student feedback into the plan-teach-reflect-cycle shifts the focus to information gathering in the facilitation of personal development of ePCK and pPCK (Elstad et al., 2017), particularly given that a teacher's behaviour in the classroom directly affects the quality of teaching (Seidel & Shavelson, 2007; Seidel & Stürmer, 2014).

Students possess the key to evaluating overall teaching quality by feedback in a cyclic process, a process that is self-enhancing and postulated in the RCM model (Wisniewski et al., 2019); this information is gathered via student ratings, which are reliable indicators of teaching quality (Kämpfe, 2009), and make efficient use of time and resources (Bijlsma et al., 2019). Building on this efficiency, mean student rating values can serve as a viable and valid measure of teaching quality (van der Scheer et al., 2019) and play an important catalytic role in the development of ePCK and pPCK (Figure 9) (Shavelson, 2020). Focusing on student ratings emphasises the idiosyncrasy of ePCK and detaches teacher skill development from the collective towards best practice individual teaching practices. The adaptation to the continuous change of student ratings leads individual teacher to focus on what is working right now and thus enables development of ePCK in practice.

Feedback tools differ significantly in their design and the way retrieved information is visualised, ranging from simple textual data to more complex graphical representations. A key aspect of effective visualisation is ensuring both usability and understandability. The role of these design principles lies in enhancing the accessibility and interpretability of visual data (Douville et al., 2025). Existing student feedback tools often rely on bar graphs and line graphs to represent various facets of teaching skills, providing detailed but static snapshots of teacher performance (Bijlsma et al., 2019; Wisniewski & Zierer, 2021). However, a shift towards focusing on the development of teaching skills over time demands

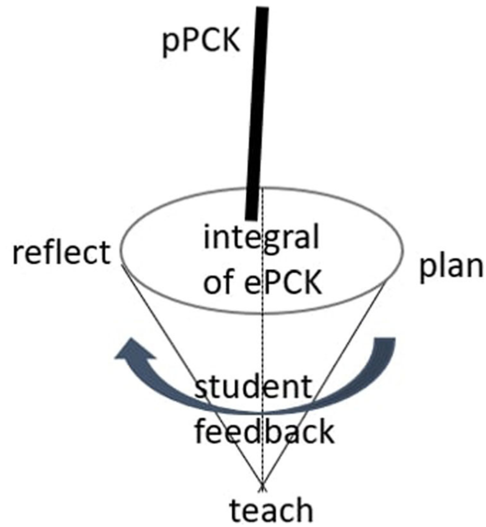


Figure 9. Reinforcing cycle of student feedback. With feedback as a catalyst, the teacher is able to use the plan-teach-reflect-cycle as a feedback loop to develop ePCK and pPCK.

visualisation methods that capture temporal trends and relative improvements. The scientific literature highlights two visualisation methods as particularly effective for this purpose: line graphs and rotation graphs (Schwabish, 2021). Line graphs are widely recognised for their simplicity and effectiveness in displaying time-dependent trends (Kubina et al., 2017), while rotation graphs offer a multidimensional perspective by illustrating relative performance and momentum over time (de Kempenaer, 2017).

This study builds on these principles by offering a tool with two distinct visualisation options tailored to support the ongoing development of teaching skills. The first option, a linear performance graph, provides a clear and straightforward representation of trends over time, making it ideal for tracking gradual improvements (Rollett et al., 2021). The second option, a rotation graph, delivers a more nuanced view by incorporating relative performance and momentum. This allows teachers not only to identify patterns and relationships in their feedback data but also to gain insights into how consistent students are in their evaluations and how quickly these evaluations change over time. Investigating which of these two visualisation approaches better supports teachers in improving their ePCK will provide actionable insights for optimising feedback tools and their implementation.

Research indicates that student evaluations tend to exhibit consistency across different courses taught by the same instructor and over a given period of time (Marsh & Hocevar, 1984). This suggests a potential correlation in student feedback across various lessons within the same course. It has been shown that student evaluations of teaching effectiveness are highly reliable and stable across different courses and instructors, indicating that students' perceptions are consistent over time (Marsh, 1984). More recently, studies have shown that instructor-related factors significantly influence evaluation scores, highlighting a level of consistency in how students rate the same instructor across different courses (Spooren et al., 2013). Furthermore, a meta-analysis of intervention studies examining the effects of student feedback on teaching suggests that feedback can lead to changes in teaching behaviour, which may, in turn, affect student evaluations in subsequent lessons (Röhl, 2021). This dynamic points to a potential dependency in ratings within a course, as student evaluations are shaped not only by static teaching characteristics but also by the evolving classroom environment.

These findings imply that student ratings may not be entirely independent but could be influenced by consistent teaching styles, instructor characteristics, classroom dynamics, or adjustments made in response to feedback. Understanding these interdependencies is crucial for interpreting aggregated feedback data accurately and enhancing the reliability and validity of such assessments over time.

Exploring whether such dependencies exist and how they affect aggregated feedback data is crucial for understanding the reliability and validity of such assessments over time (Kahneman et al., 2021; Wisniewski & Zierer, 2021).

Research objectives and research questions

Integrating regular student feedback into the routines of the classroom environment could be an effective way of implementing a feedback process used in a teacher's plan-teach-reflect-cycle. This could help them systematically enhance their ePCK and teaching strategies by providing the required empirical data.

The main goal of the present study was to find out if certain visual representations of student feedback for teachers are effective in fostering ePCK of the RCM Model.

We considered the following research questions:

RQ1: Can a student evaluation tool for continuous lesson rating be developed and implemented in a way that improves an individual teacher's ePCK over time via student feedback?
RQ2: Which of two visual options helps teachers more in improving ePCK over time?
RQ3: Are student ratings of various lessons within one course dependent on each other?

Methods

Study design

The current work was an intervention study. In the week before the study period, from 17 April to 24 April 2023, the teachers were familiarised with the feedback tool and its two types of visualisation. The study commenced on 25 April 2023 and concluded five weeks later at the beginning of the Pentecost holidays in Germany, of the same year. The participating teachers were randomly assigned to three groups (Groups A, B, C), each comprising eight to nine teachers. All groups started with no visualisation in the first week (since there were no data to be depicted) with visualisations introduced from the second week. As May 1 was a national holiday, the second week started at Tuesday May 2 and all consecutive study weeks were shifted by a day so that they also began on a Tuesday. In weeks three to five, each group had a week of Performance (results shown as a moving average curve (y axis) of the student ratings over time (x-axis)), Fluctuation (results shown as a two-dimensional counter-clockwise curve of the average student ratings (y-axis) and the slope of the average student ratings (x-axis) and no visual presentation, but in different order. May 26 was the end of the study and the last school day before the holidays. No lessons were held on May 18 as it was a national holiday. Table 1 shows the study design for all groups.

Instruments

The student rating environment Teacher Tunes

The study was conducted by using the web-based feedback environment *Teacher Tunes* (teachertunes.de) which was developed at the LMU Munich. The tool collects student feedback of a lesson via a rating system. Each teacher registers for the tool with his/her own account and generates separate courses for each participating class. After each lesson, students rate the lesson with the help of their smartphones and a QR-Code on a five-star rating scale. The student feedback for teachers was visualised by a curve of the moving average (of the three most recent lessons) of the student rating and shown to the teachers on their dashboard. Depending on their assigned treatment group teachers saw one or none of three visual graphics of the student feedback on their *Teacher Tunes* dashboard: (1) none, (2) performance (see Figure 10) and (3) fluctuation (see Figure 13(a)).

Table 1. Study design.

Design group	Timeline Start date	Introduction 17 April 2023	Week 1 24 April 2023	Week 2 02 May 2023	Week 3 09 May 2023	Week 4 16 May 2023	Week 5 23 May 2023
A		–	None	Performance	Performance	None	Fluctuation
B		–	None	Performance	None	Fluctuation	Performance
C		–	None	Performance	Fluctuation	Performance	None

Notes: The three groups of teachers over the five weeks of the study and the initial introduction. Each group worked with different visualisation types in different weeks (see Figure 18).

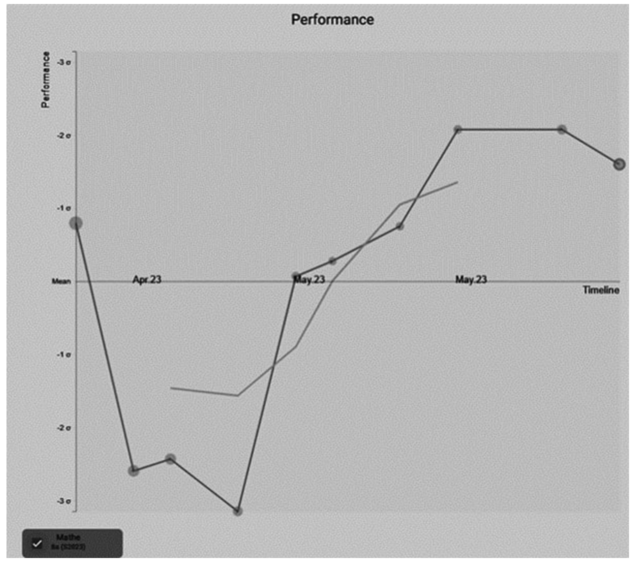


Figure 10. The bold visualisation type performance moving average curve. The curve (MA = 3; average of the three latest lesson ratings) shows the development of the student ratings (y-axis) of a certain class over time (x-axis).

To ensure data reliability, especially for student feedback collected via smartphones, the rating system was designed with robust mechanisms to guarantee fair participation and accuracy. The tool centres around a feedback mechanism where students provide ratings for specific lessons. Each teacher was assigned a unique QR code. By scanning this code, students were directly redirected to the relevant lesson's rating page. Alternatively, students could manually input the course name and a unique lesson PIN to access the rating interface, ensuring that every student had the opportunity to rate each lesson regardless of technical constraints or device compatibility.

A strict 'one rating per student per lesson' rule was implemented to maintain the integrity of the data. When a student submitted a rating, the system generated a unique token that was stored locally on the student's browser. This token acted as an identifier to prevent multiple submissions for the same lesson. If a student attempted to rate the same lesson again, the system replaced the previous rating with the new one, effectively ensuring that only the most recent input was recorded while still adhering to the single-rating policy. This approach not only prevented duplicate entries but also allowed students to update their feedback if necessary.

The system was designed to respect student anonymity and minimise data collection. Only the token, the rating and the date of submission were stored, providing sufficient data for analysis while safeguarding privacy. Each rating was tied to the specific lesson within the respective course, ensuring that feedback was accurately attributed. The user interface was kept minimalist and intuitive, with easy access to the course and lesson identification features. Upon submitting a rating, students received a confirmation message, reassuring them that their feedback had been successfully recorded.

Design of the graphic visualisations

The visualisation type 'performance' displays in a bold line the moving average of the three latest lesson ratings (MA = 3) of the student ratings in a given class (y-axis) over the timespan of the study (x-axis) (Figure 10). The absolute values of the average student ratings are not displayed but the relative average student ratings to the total variance of the student ratings are shown. The zero point of the y-axis is the arithmetic mean of all courses of a specific teacher. The diameter of the bold circles shows the total number of students contributing to that datapoint. The more student ratings that contribute to one data point the larger the diameter of the respective circle. The fine grey line shows the moving average curve of all courses of a specific teacher.

Visualisation type 'performance' shows the relative performance curve of more than one class of a specific teacher. Thus, student ratings of different classes of one teacher can be compared visually (Figure 11).

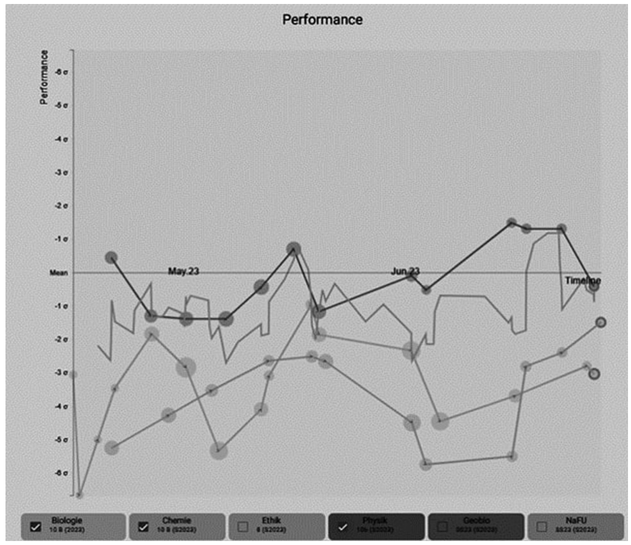


Figure 11. Overview of several courses. Three classes (biology, chemistry and physics) given by a teacher are displayed. The courses biology and chemistry receive bad student ratings relative to all other classes of the teacher.

As a benchmark for comparison, two options are available. In the default setting of the tool *Teacher Tunes*, the zero point of the y-axis is the arithmetic? mean of all courses of a specific teacher, which generates an intrapersonal benchmark. In the alternative setting, the zero point of the y-axis is the arithmetic? mean of all courses of all participating teachers in the study, which generates an interpersonal benchmark. Teachers can switch between these options by clicking the relevant button on *Teacher Tunes* as illustrated in Figure 12.

The visualisation type Fluctuation displays in a bold line a moving average (MA = 3) of the average student ratings in a given class (y-axis) against the slope of the relative moving average (MA = 3) (x-axis). The steeper the gradient, the farther away from the crosshair the point is depicted on the x-axis. If the slope of a datapoint is positive, the relative moving average curve is depicted on the left-hand side of the crosshair and vice versa. The small arrow in the bold circles shows the unfolding of the curve over time. Thus, the development of the student ratings shows a counterclockwise movement going through all four quadrants. Ideally, in the case of a sinusoidal curve, the visualisation depicts a symmetrical circle with the mean student rating at the centre of the circle. According to this counterclockwise movement, the quadrants have been labelled to describe the respective development of a sinus curve. The ideal (sinusoid curve) unfolding and the respective labelling of the quadrants is shown in Figure 13.

Teachers can switch between these options with a click on the respective button on *Teacher Tunes* as is illustrated in Figure 14.

Justification for the specific timeframes and intervals used in student feedback visualisations

The selection of specific timeframes and intervals, such as the three-lesson moving average curves and the aggregation approach used in our study, is rooted in the typical teaching and learning dynamics observed in schools and the nature of our research objectives.

Our study aims to evaluate the development of teaching quality and its perception by students over time rather than pinpointing feedback tied to specific lessons. The chosen timeframe and intervals allow us to smooth out short-term fluctuations in feedback that might be influenced by external or transient factors (eg unique events during a single lesson) and instead highlight meaningful trends or shifts (Kahneman et al., 2021)

In most school settings, science teachers conduct two to four lessons per week for a given class, and this frequency suggests that analysing individual lessons in isolation may not adequately capture patterns or trends in teaching effectiveness based on student feedback. Aggregating feedback over slightly longer intervals aligns better with the pace of classroom interactions and the rhythm of instructional delivery (LeBreton et al., 2023).

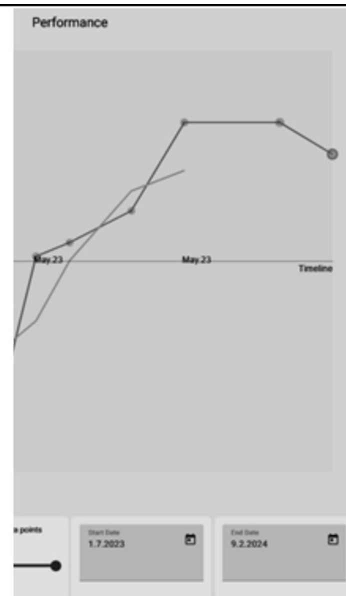
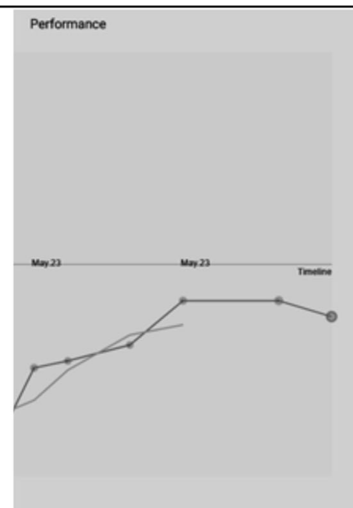
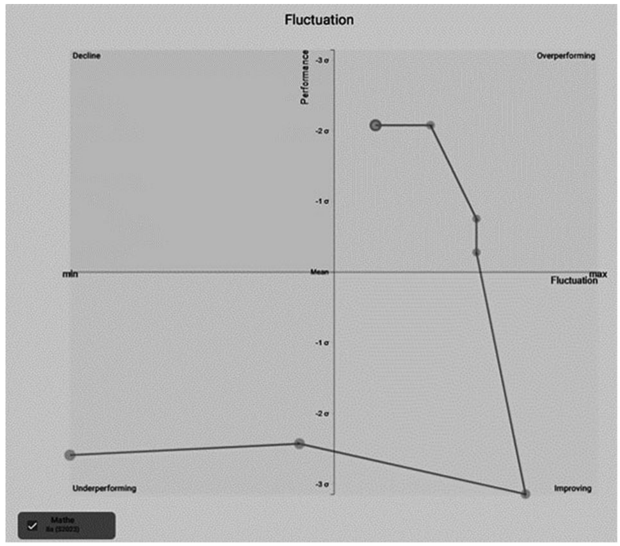
Benchmark: Comparison Intrapersonal The zero point of the y -axis is th arithmetic mean of all courses a specific teacher	
Benchmark: Comparison Interpersonal The zero point of the y -axis is th arithmetic mean of all courses all teachers in the study	

Figure 12. Benchmark options. Each teacher can choose either benchmark option. The intrapersonal benchmark compares the performance curve of a class to the average of all performance ratings of the teacher. The interpersonal benchmark compares the performance curve of a class to the average of all performance ratings of all teachers.

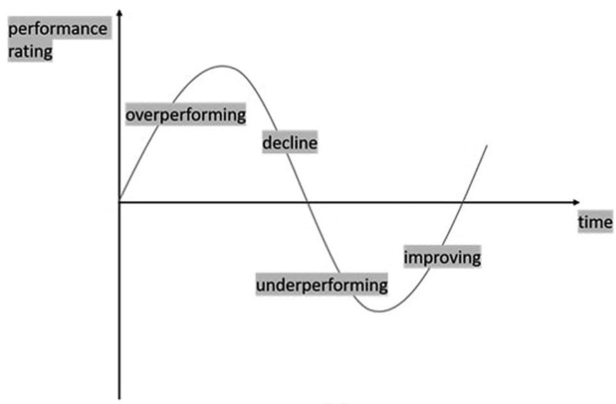
Typical topics or concepts in school are often taught over multiple lessons, and aggregating feedback over intervals longer than a single lesson ensures that the analysis accounts for the entire instructional arc of a topic. This approach provides the option of a more holistic view of how students perceived the teaching of a concept or theme.

Improvements or declines in teaching quality are often expected to manifest within one to two weeks, given the cumulative nature of student learning and perception. Aggregating data over this interval ensures that both gradual and more pronounced changes are visible, without introducing excessive lag that might obscure shorter-term effects (Gaertner & Brunner, 2018).

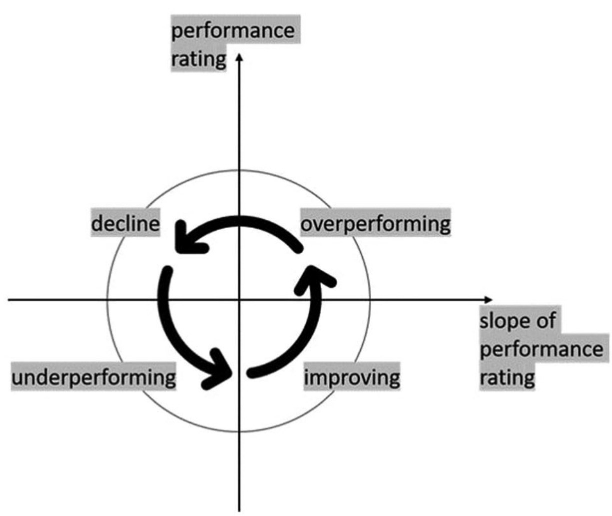
Furthermore, the five-week duration of our study necessitated a data aggregation mechanism that is proportionate to this period. Using a three-lesson moving average allowed us to balance the need for trend clarity with the available timeframe, ensuring that any observed effects could still be meaningfully attributed to the intervention or changes in teaching methods during the study.



(a)



(b)



(c)

Figure 13. (a) The bold visualisation type fluctuation moving average curve. A typical counterclockwise moving visualisation of student ratings. (b) Idealised performance curve over time. The quadrants of the visualisation fluctuation map to the four stages of a sinusoidal performance curve (overperforming, decline, underperforming, improving). (c) Counterclockwise rotation. An idealised visualisation fluctuation depicts a cyclical counter-clockwise rotation. The bold visualisation type fluctuation moving average curve. The development of a teacher's performance rating over time can be illustrated as a counterclockwise movement. The four quadrants show the state of performance rating development as overperforming, decline, underperforming and improving.

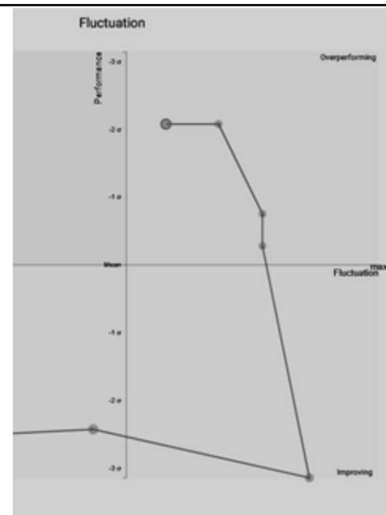
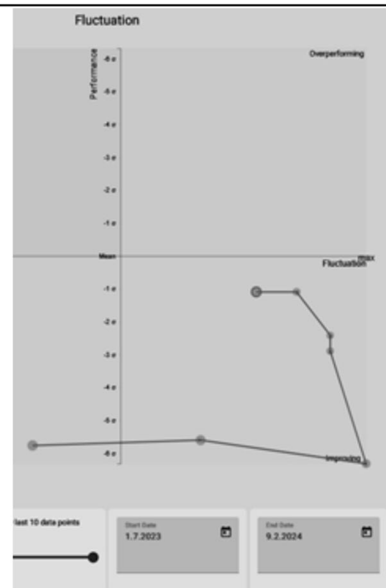
Benchmark: Comparison Intrapersonal *The zero point of the x* the y-axis is th arithmetic mean of all courses o specific teacher *-axis is th firstderivative of the y-axis of al courses of a specific teacher*	
Benchmark: Comparison Interpersonal *The zero point of the x* -axis is th arithmetic mean of all courses all teachers in the study *-axis is th first derivative of the y-axis ofall courses of all teachers in the*	

Figure 14. Benchmark options. Each teacher can choose between two benchmark options. The intrapersonal benchmark compares the performance curve and slope of a class to the average of all performance ratings and to the slope of these ratings of the teacher. The interpersonal benchmark compares the performance curve of a class to the average of all performance ratings of all teachers and to the slope of these ratings of the teachers.

Furthermore, the five-week duration of our study, combined with an additional onboarding week for the teachers, aligns well with the theory by covering the period between the Easter and Pentecost holidays. This timing increases the likelihood that the lessons reflect a coherent thematic structure or focus on a single subject area, ensuring content continuity. Using a three-lesson moving average allowed us to balance the need for trend clarity with the available timeframe, ensuring that any observed effects could still be meaningfully attributed to interventions of teachers and changes in teaching methods during the study.

Data preparation

Sample

The sample comprised 28 secondary school science teachers (52% females, 48% males). The majority of the participating teachers were from 20 to 37 years old, with one teacher being 57 years old (average age 34, SD = 7.41, Min = 20, Max = 57). Each science teacher took part in at least one class with a total of 48 courses. In total, data of 199 lessons from 535 students were collected.

Table 2. Number of lessons rated in the five weeks of the study.

Week	1	2	3	4	5
Number of lessons	39	36	34	18	10

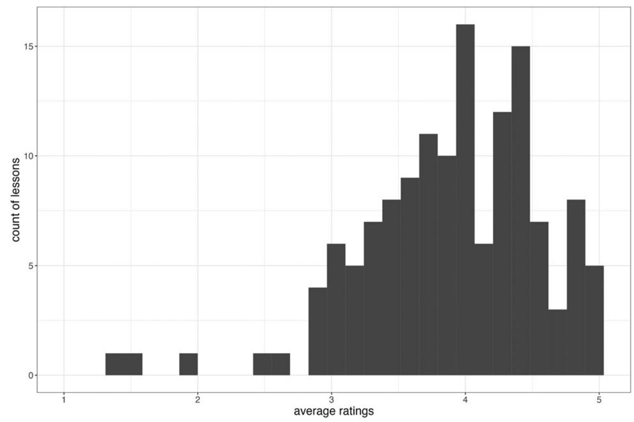


Figure 15. Histogram of the average student ratings. Most fall around 4.

Data cleaning

The data were downloaded from the *Teacher Tunes* website. After adding the teachers' age and sex, the data were anonymised. Each teacher-class-combination was merged to form a new randomly generated id.

All lessons with zero ratings were excluded. Since this study was about the change of lesson ratings over time, the id's (classes) that had only one rated lesson were removed. All lectures with fewer than five student ratings were excluded to prevent single student impressions affecting the inference. The regression model used to quantify the association strength between average rating and the visualisation generated (see following sections) could not deal with missing values and therefore 15 more lessons where age and sex were missing, were excluded, leaving 137 lessons for analysis.

Data description

The sample of 137 consisted of 28 unique teacher-class combinations that were rated between two and 12 times; 67 of the lessons were held by a female teacher, 70 by a male. Most of the teachers were from 24 to 37 years old (average age =35. SD = 6.62) with one teacher of 57. Table 2 shows the number of lessons per week rated by student feedback. During the first three weeks more, lessons were rated than during the last two weeks of the study.

Figure 15 shows the average student ratings. Although the feedback scale ranged from 1 to 5, 93.4% (128) of them lay above 3 with a mean of 3.9.

The number of ratings lay from 5 (see data cleaning) to 24. The lessons were observed between April 24 and May 26, except on the holidays May 1 and 18. Figure 16 shows the number of ratings (left plot) and the average rating (right plot) over the study period. A point relates to a lesson and the darker the point the more lessons had this value. For neither the number of ratings nor the average rating can a change over time be seen.

The participating teachers were randomised into three treatment groups (A, B, C). Table 3 shows the number of rated lessons for each group.

The distribution of data in the visualisation on the day of the lecture is shown in Table 4. One scenario is that at some point on the day of the lecture, the teacher checks the feedback tool website and uses this information for lecture preparation. However, during lecture preparation teachers received no visual feedback 60 times, observed Performance 59 times and the Fluctuation 18 times.

Figure 17 depicts the seen number of website visits and number of lectures over the whole study period.

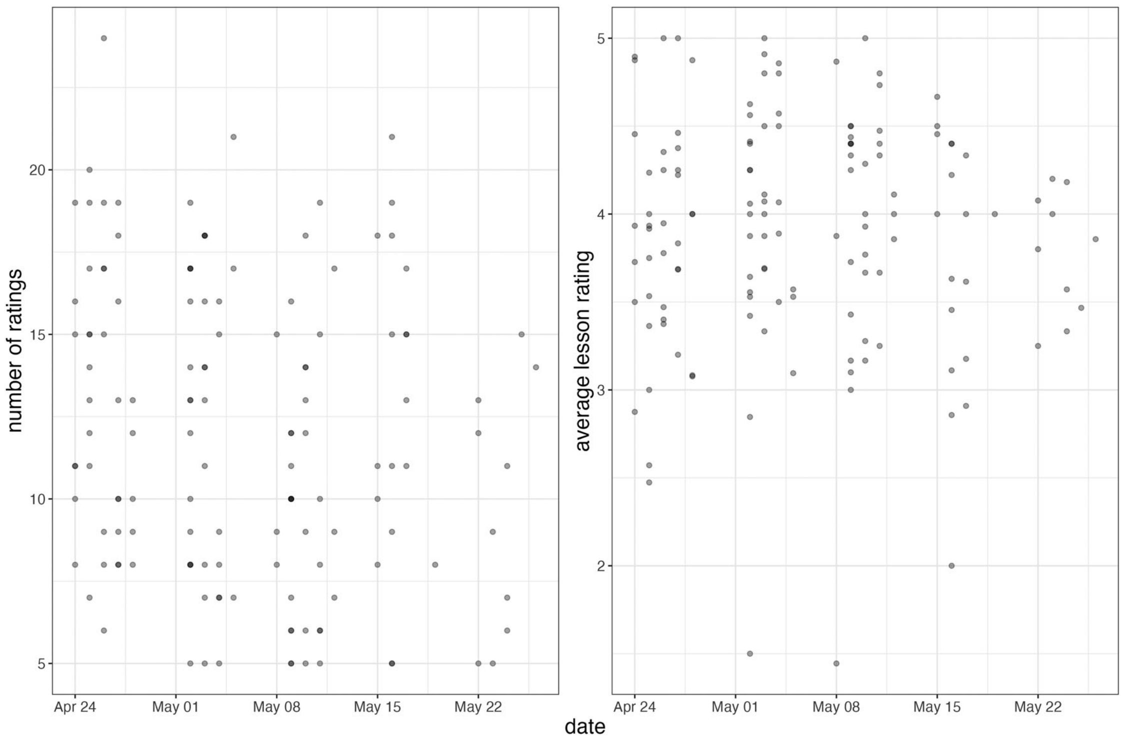


Figure 16. Lesson date (x-axis) and number of ratings (left) and average ratings (right). Neither the number of ratings nor the average rating change over time.

Table 3. Number of rated lessons in the treatment groups.

Treatment group	A	B	C
Number of lessons	65	46	26

Note: From a total of 137 rated lessons, 65 were taught by teachers in group A, 46 in group B and 26 in group C.

Table 4. Number of rated lessons where the teacher checked a visualisation on the same day.

Visualisation observed (day of lecture)	None	Performance	Fluctuation
Number of lessons	60	59	18

Table 5 shows that the mean of the average ratings for Performance (4.01) was higher than for Fluctuation (3.62) or no visualisation (3.86). The standard deviation (SD) and ratio of the number of ratings per rated lesson are similar.

One scenario is that teachers take the option to check their history of student ratings for reflection and planning purposes the day before the lecture, especially if it is early in the day. The type of visualisation referred to on the day before the lecture is shown in Table 6. Shifting the type of visualisation by a day changes the type of visualisations only on the starting days of the intervals, ie the Tuesdays.

Figure 18 depicts the type of visualisation accessed on the day before the lecture together over the study period.

Regression model

A regression model was used to estimate the strength of association between the average rating and the type of visualisation, with average rating as the dependent variable and type of visualisation as the independent. The model controls for the teachers' age, sex and the lesson date. The number of ratings is associated with the variance of the dependent variable and therefore cannot be included as a confounder.

As different classes have different performance levels and lectures of the same class are likely to be correlated (see research question 3 and section Evaluation of research question 3), these factors were

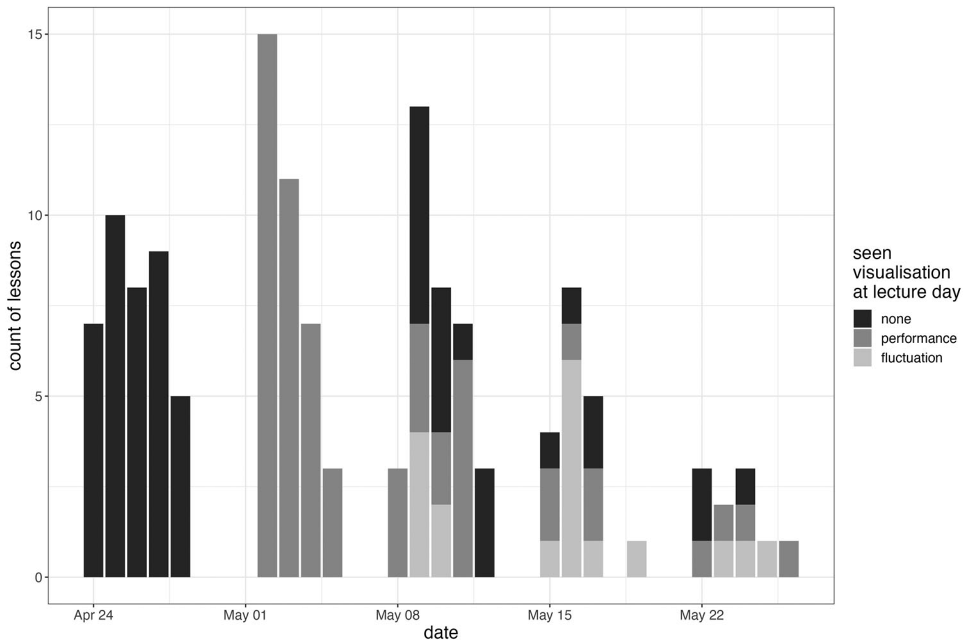


Figure 17. Distribution of the observed visualisations. The number of lectures held (y-axis) over the study period (x-axis) shaded by the type of visualisation observed on the day of the lecture is illustrated.

Table 5. Overview statistics for the visualisations.

Visualisation type	Number of lessons	Average number of ratings per lesson	Mean average rating	Standard deviation
None	60	12.5	3.86	0.61
Performance	59	11.6	4.01	0.69
Fluctuation	18	10.6	3.62	0.67

Note: The mean average rating is highest for visualisation performance.

Table 6. Number of rated lessons where the teacher referred to a certain visualisation on the day before the lecture.

Visualisation observed (day before lecture)	None	Performance	Fluctuation
Number of lessons	75	53	9

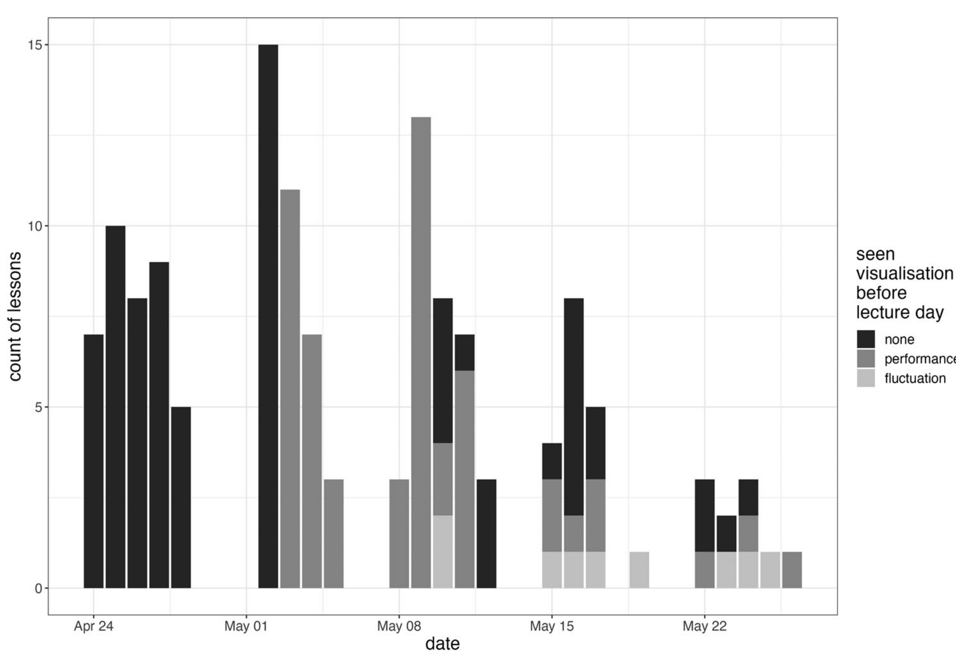


Figure 18. Distribution of the type of visualisation accessed before lecture day. The number of held lectures (y-axis) over the study period (x-axis) shaded by the type of visualisation.

considered in the model by extending it by a class-specific random intercept. This linear mixed model (Fahrmeir et al., 2013) was estimated with the open source software R and its package lme4 (Bates et al., 2015).

Teachers could access the feedback tool at any time so could use the visualisation during a lecture, integrating it into a feedback discussion with their students, or during the preparation time the day before to reflect on previous student feedback and perhaps prepare the lecture differently, but the teachers' preparation behaviour was not surveyed. It is further possible that the analysis may be missing information from potential further confounders, such as other sources of feedback or time to the next graded test. Even if more information were available, the present sample size does not allow for arbitrarily complex models, in order to obtain meaningful uncertainty estimates and avoid numerical instabilities. Besides the complexity, the explanatory power is also limited by the sample size. This limitation is particularly pertinent for the fluctuation type of visualisation, as there were only 18 lectures that were prepared with the fluctuation type of visualisation on lecture day and even fewer on the day before.

In the next section, we report the results for two models: with the type of visualisation accessed on the day of the lecture versus on the day before.

Results

Interpretation of coefficients of regression

The regression coefficients for the model with feedback accessed on the day of the lecture together with their standard errors are shown in Table 7.

The intercept cannot be interpreted meaningfully and is only of secondary interest. The lesson date and teachers' age and sex are included for their possible confounding effects and are also only of secondary interest. The estimated associations are interpreted in the following manner: if a lecture in the same class is held a day later and the other variables are fixed the average rating increases on average by 0.0036. If a lecture in the same class is held by a teacher who is a year older, the average rating increases on average by 0.0003. If a lecture in the same class is held by a male rather than a female teacher, the average rating increases on average by 0.0080. The standard errors of lesson date, and especially teachers' age and sex are (much) higher than the coefficients themselves, so the model does not provide clear evidence.

The main goal of this model was to investigate the effectiveness of visual representations of student feedback for teachers in fostering ePCK. In statistical terms, we were interested in quantifying the strength of association between the average rating and the accessed visualisation. If a lecture for the same class had been held by a teacher who had received feedback in form of Performance, the average rating increased by 0.0686 (SE = 0.1059) compared with a teacher who received no such graphic feedback. Contrary to expectations, feedback in the form of Fluctuation decreased the average rating by 0.2388 (SE = 0.1691) compared with no visualisation. However, the results have to be assessed with some caution as the standard errors for the two coefficients are close to the mean values and the model had only 18 observations for the fluctuation type of visualisation.

Table 8 shows the regression coefficients for the model with the feedback accessed on the day before the lecture. For the intercept, lesson date, age and sex, the only difference from the previous model was

Table 7. Model regression coefficients when feedback was accessed on the lecture day.

Variable name	Intercept	Lesson date	Age	Sex – male	Seen vis. – performance	Seen vis. – fluctuation
Estimated coefficient	3.9282	0.0036	0.0003	0.0080	0.0686	−0.2388
Standard error	0.4875	0.0065	0.0131	0.1810	0.1059	0.1691

Table 8. Regression coefficients with visualisation accessed the day before the lecture.

Variable	Intercept	Lesson date	Age	Sex – male	Seen vis. – performance	Seen vis. – fluctuation
Estimated coefficient	3.9247	−0.0034	0.0010	0.0012	0.1670	0.0927
Standard error	0.4841	0.0064	0.0130	0.1817	0.1037	0.2187

the opposite sign for the lesson date (the later a lesson is held during the study period the lower the average rating).

If a lecture in the same class was held by a teacher who had received performance type feedback during lecture preparation the day before it was given, the average rating increased by 0.1670 (SE = 0.1037) compared with a teacher who received no graphic feedback. There were only nine lectures where the teacher received feedback via Fluctuation the day before the lecture. The average increase of the average rating by 0.0927 (SE = 0.2187) has to be evaluated with caution.

The two models were also evaluated by non-linear associations for lesson date and age, and in an analysis of sensitivity, the models were re-estimated excluding data from the 57-year-old teacher. In all these scenarios, the tendencies and signs of the visualisation coefficients were unchanged.

In summary, the Performance did not significantly improve the average rating. To evaluate the Fluctuation thoroughly, more data are needed.

Evaluation of research questions

Evaluation of research question 1

Research question 1 was whether usage of the feedback tool could improve teacher ePCK as measured by student rating over time. The regression model revealed a positive tendency for the student rating for Performance to improve when it was used both during preparation and on the lecture day. For Fluctuation the tendency was not clear and dependent on the time of use. To assess if the tool improved the student rating significantly ($p = .05$), a likelihood-ratio test compared models with and without the visualisation information. The null hypothesis is that the estimated associations of the two types of visualisation are zero (the visualisation has no connection with the average rating). From the sample size and standard errors, the authors did not expect a clear rejection of the null hypothesis.

For the models with visualisation on the lecture day, the *p* value was $p = .1429$ and with visualisation on the previous day it was $p = .2753$. Consequently, the null hypotheses could not be rejected. In this case, the *p* values are rather small and near to the 0.05 threshold, suggesting the possibility of a connection of the visualisations with the average rating. Nevertheless there is not enough evidence in the sample for a definitive conclusion.

Evaluation of research question 2

The second goal of the study was to compare performance with the fluctuation type of visualisation. As evaluation of the fluctuation types was limited by its low number of observations (18 on the lecture day, 9 on the previous day), a likelihood ratio test was used, in which both models were compared with a model with only a binary indicator (whether a person viewed or did not view any visualisation). This corresponds to the null hypothesis that the association strength for Fluctuation and Performance is the same; the authors did not expect a clear rejection of the null hypothesis.

For the model for the visualisation on the lecture day $p = .1121$ and for the previous day $p = .7211$, so the null hypotheses could not be rejected. Notably, despite the small number of observations, the *p* value of .1121 is remarkably low, suggesting that there may be a difference in the way the different types of visualisation affect the average rating when viewed on the day of the lecture. The limited number of cases in which the fluctuation type of visualisation was observed on the previous day ($n = 9$) introduces considerable uncertainty, and the result must be interpreted with caution, as it is possible that the data do not reveal the true underlying pattern. For a thorough analysis a more extensive dataset would be required.

Table 9. Regression analysis.

	RI variance	Residual variance	*p* Value	ICC
Model with vis. on lecture day	0.1552	0.2667	>.0001	0.360
Model with vis. on previous day	0.1550	0.2697	>.0001	0.360

Note: In both model cases, *p* values are close to zero and thus the corresponding null-hypotheses are rejected.

Evaluation of research question 3

Research question 3 was about the correlation of multiple lessons in the same class.

For regression analysis the structure of the data must be known. It was assumed that the responses to/feedback on? lessons of a teacher in the same class were correlated and the model was accordingly adjusted with a class-specific random intercept (RI). The variance of this random term is estimated and shown for both model versions in the first column of Table 9. The random effect variance is in both cases about 0.155. The hypothesis that the variance of the random intercept is equal to zero, and therefore that the random intercept is not needed, was tested by the R-package RLRs in following the methodology of Scheipl et al. (2008). In both cases, the *p* values were very close to 0 (shown in the third column of Table 9) and thus the null hypothesis was rejected.

From the random intercept variance and the residual (the remaining unexplained variance of the models), the intra-class correlation coefficient (ICC) can be computed as random intercept variance divided by the total variance (the sum of both variances). The values are shown in column four of Table 9 and both are about 0.360, which implies that the average ratings of two lessons of a teacher in the same class are positively correlated.

Discussion and conclusion

The aim of this study was to investigate whether visualisations of student feedback on teachers' performance in class are effective in improving ePCK of the RCM Model.

Influence of student feedback on ePCK

Our feedback tool for student rating delivered promising results. Student feedback helped teachers enhance their ePCK as it was associated with mitigation of the student assessment of lesson quality during both the teaching phase and preparation and planning phase of the cycle (Wisniewski & Zierer, 2021). Considered as one slice of ePCK a teacher shows in a given lesson, and assessed via student rating, what makes the top spin is bringing this information into the next performance, during the next lesson, with a different topic, helping a teacher to develop ePCK over time. Using critical information on past performance to improve ePCK in future performance situations in a dynamic teaching environment fulfils the demand for cyclical improvement entailed and generated through the evolution of the RCM. Integrating such reinforcing feedback loops into the process of teaching, by means of a visualisation of student feedback, successfully connects the two iterative processes of the teaching cycle of the RCM model and the feedback cycle, thus generating a possibly reinforcing cycle structure (Carlson et al., 2019).

The question of what constitutes critical information for improving ePCK remains not fully answered. Following the simple reasoning that relative data are more robust than absolute values of ePCK, which is difficult to measure, we focused on the use of relative change as valuable information (Taleb, 2012). But whether relative performance, rather than absolute values, is the right kind of information to use in a reinforcing cycle may not only depend on the dynamics of the environment but also on individual preferences and needs. For example, a teacher might simply want to know whether he is doing satisfactorily, whilst not having the resources to focus on adapting in an ever-changing environment.

Comparison of the two different types of visualisation

The Performance visualisation yielded a clearer and stronger positive influence on the development of teacher ePCK compared to Fluctuation. This difference highlights the importance of simplicity in data representation, particularly when teachers must quickly interpret feedback and integrate it into their teaching practice. While the Fluctuation visualisation offers a more detailed depiction, its complexity introduces several challenges that limit its effectiveness. Below, key factors contributing to this discrepancy are discussed, along with potential improvements.

The Fluctuation visualisation is less intuitive than the Performance representation because interpreting it requires teachers to simultaneously consider two dimensions: the student ratings on the y-axis and the momentum – how quickly ratings change and in which direction. This dual focus significantly increases cognitive load, making it harder for teachers to derive actionable insights (Guo & Chen, 2023). For teachers already balancing multiple responsibilities, this added complexity diminishes the usability of the Fluctuation visualisation.

Unlike the straightforward line graph in the Performance visualisation, the two-dimensional crosshair design of Fluctuation can obscure clear trends. The simplicity of the Performance visualisation provides a clear trajectory, making trends over time easy to identify. In contrast, the interaction between dimensions in the Fluctuation visualisation may complicate interpretation, hindering quick and informed decision-making (Mueller et al., 2024). Simpler visual formats, such as line graphs, are more effective for identifying trends as they better align with users' mental models and expectations.

The momentum dimension in the Fluctuation visualisation may also limit its actionability. While it illustrates the rate and direction of change, it can feel abstract and disconnected from practical classroom adjustments. In comparison, the Performance visualisation offers direct insights into student perception trends, enabling teachers to evaluate lesson effectiveness more easily (Wisniewski et al., 2019). Feedback mechanisms are most effective when they provide clear and actionable insights, which is more readily achieved with the simplicity of the Performance visualisation.

Another contributing factor is familiarity. Line graphs are commonly used and widely understood in educational settings, requiring little to no additional training to interpret. Conversely, the Fluctuation visualisation's unfamiliar format may discourage its use or lead to misinterpretation. Users tend to perform better with familiar visualisation styles, as these reduce cognitive barriers and foster trust in the data (Viberg et al., 2018). Without proper training, teachers may struggle to effectively utilise the benefits of the Fluctuation visualisation.

One potential improvement to the Fluctuation visualisation would be to replace momentum with a measure of student consensus, such as the variance of student ratings within a class. This approach would highlight whether the class agrees on lesson quality or if significant disparities exist. For instance, high variance might suggest the need for differentiation or indicate that some students are disengaged, while low variance would reflect a shared perception of lesson quality. This adjustment could provide clearer and more actionable insights, directly tied to classroom dynamics. Further research and testing would be required to validate this approach (Röhl, 2021).

The extended duration of the five-week study likely exacerbated these challenges, as participant fatigue led to a notable decrease in rated lessons towards the end of the study (Döring & Bortz, 2016). This underscores the importance of designing visualisations that minimise cognitive demands and maximise usability to ensure sustained engagement and effective feedback utilisation (Döring & Bortz, 2016; Moosbrugger & Kelava, 2012; Taleb, 2012).

Usage of within-class rolling averages for feedback visualisation

Our data clearly suggest that for one class and one specific teacher, consecutive lesson ratings are highly correlated, which indicates that student ratings of a specific course can be interpreted as a continuous process. To underscore this process and focus on the development of student ratings over time, the use of rolling averages for the feedback visualisation has proved to be apt (Kahneman et al., 2021). Developing ePCK within the one teacher-one class relationship might constitute a dyadic partnership that forms the fruitful teaching context postulated in the RCM – the right kind of environment for a teacher to improve his ePCK. The classroom environment or learning context of a given class serves as the stable environment for the spinning top of ePCK to unfold with the help of student feedback.

Limitations

In this study the number of participants was low. With more teachers instructed to use the feedback tool in only one specific way, or in one specific part of the plan-teach-reflect-cycle, variance would be reduced and the chance of providing valid conclusions augmented.

Despite the small sample size, which can affect the validity of the results (Döring & Bortz, 2016), we achieved reliable results for research question three and strong but not statistically reliable tendencies for research questions one and two, and thus can derive generalisations. Implementing a new visualisation tool into the teaching cycle is a learning process in itself that takes time and getting used to Shavelson (2020). From studies on habit formation, approximately two months are needed to acquire a new one. Using a visualisation tool as a cyclical process to foster lesson development and develop ePCK and pPCK might have needed more time than that in our study (J. Keller et al., 2021).

Several studies that focus on interventions in real teaching situations assess student grades or content-specific task results (Mientus et al., 2022). A possible link between student achievement and their learning progress with student feedback for teachers is not tackled by this study and remains unclear (Shavelson, 2020).

The focus on Individual development disregards the effect of personal PCK development on that of collective PCK of the community. This might not be so in professional communities of teachers but is not further investigated in this study (Altrichter et al., 2018).

Integration of our results and interpretations into the Refined Consensus Model of PCK

In comparison with other studies on the RCM, the most distinct feature of our study is its focus on integrating student feedback for teachers into their plan-teach-reflect-cycle, thus reinforcing the feedback-loop for teacher and improving teaching (Alonzo et al., 2019).

As postulated by the RCM the visualisation of student feedback for teachers focuses on the iterative process of an individual's development (Mientus et al., 2022).

From our data, student feedback can be successfully integrated during the planning and reflecting phase, and the teaching phase of the teaching cycle, enabling interaction with students (Wisniewski et al., 2019), empowering them to shape lesson development and the strategies of their teacher (Bijlsma et al., 2021). However, the data gathered over the five-week course did not provide definitive conclusions from the teachers (Bijlsma et al., 2019)

The RCM is to provide the means for dynamic and individual development, and the feasibility shown here of a simple student rating allows the implementation of a feedback cycle in the teachers' everyday teaching routine.

These discoveries support the enhancement of reciprocal feedback systems and encourage further investigation in this field.

Although the initial findings in this study were inconclusive, our data suggest the possible effectiveness of student feedback, in the form of basic graphical representations, as a means for teachers to modify their teaching quality, gradually improving ePCK over time. As such it is a profitable approach to integrate student's feedback into the RCM and to underline the dynamic nature of ePCK.

Implications for future research and teacher training programs

This study reveals several promising directions for future research and applications in teacher training programs, with a focus on implementing feedback mechanisms into a dynamic plan-teach-reflect cycle and broadening their application to diverse educational contexts.

A first direction for future research involves investigating the usability and likability of TeacherTunes. Using established usability metrics, such as the User Experience Questionnaire (UEQ) (Schrepp et al., 2014) or conducting teacher interviews could identify barriers to adoption and generate ideas for new features. For example, allowing students to provide specific comments alongside ratings could offer teachers richer, qualitative insights into lesson quality. Understanding how teachers perceive the tool's usability could guide further development and align it with their practical needs (Wisniewski et al., 2022).

Another promising avenue is refining the Fluctuation visualisation by replacing momentum with a measure of student consensus, such as the variance of ratings. High variance might indicate the need for differentiation or suggest that a significant portion of the class struggles to follow the lesson, while low variance reflects agreement in perceptions of lesson quality. Teacher feedback on this adjusted visualisation could reveal whether it better supports classroom decision-making and enhances reflective teaching practices (Rollett et al., 2021).

Further research could explore how teachers use student feedback during different phases of the plan-teach-reflect cycle. For instance, studies could investigate whether reflective use of feedback improves lesson planning or if real-time usage fosters immediate improvements in teaching (Carlson et al., 2019).

Future studies could also compare the impact of student feedback across different subjects, educational levels and age groups. For example, investigating how student feedback supports university-level teaching versus secondary school lessons could highlight its adaptability to diverse educational contexts. Subject-specific dynamics or curriculum structures might influence how effectively feedback mechanisms improve teaching quality (Behling et al., 2022).

For contexts where smartphones are unavailable (eg primary schools), alternative tools such as physical feedback buzzers could be developed. Such solutions would ensure equitable access to feedback systems and could reveal whether simpler, low-tech alternatives can match the effectiveness of digital tools. This direction aligns with findings on how digitally measured feedback improves teaching practices and could extend these benefits to less digitally equipped classrooms (Bijlsma et al., 2022).

Moreover, our student feedback tool could be employed to evaluate the impact of specific didactic interventions, such as *catch and hold* strategies or pre-planned learning progressions in parallel classes. Research could examine whether TeacherTunes can effectively measure and enhance the impact of these targeted teaching innovations, building on previous findings that feedback mechanisms play a critical role in supporting effective teaching (Shavelson, 2020; Wisniewski et al., 2019).

In addition, this study highlights actionable recommendations for integrating feedback mechanisms like TeacherTunes into teacher training programs. Embedding feedback into the plan-teach-reflect cycle is also essential. Teachers should learn to collect feedback during lessons, reflect on it and adjust their strategies. This approach supports professional growth and aligns with the RCM which emphasises the iterative relationship between reflection and practice (Carlson et al., 2019; Irmer et al., 2024).

Tailored training is vital to address the unique challenges of different subjects and educational contexts. Strategies must align with specific teaching domains, enhancing the applicability of feedback systems across diverse environments (Behling et al., 2022).

Finally, fostering collaborative feedback practices can amplify the benefits of feedback mechanisms. Encouraging teachers to engage in professional learning communities promotes shared reflection and collective improvement, strengthening instructional quality (Rollett et al., 2021).

By addressing these areas, future research could not only improve the functionality and reach of our tool TeacherTunes but also deepen our understanding of how feedback mechanisms influence teaching practices across various educational contexts. These insights have the potential to inform both practical applications in teacher training and theoretical advancements in understanding the dynamic nature of pedagogical content knowledge (Shavelson, 2023).

Ethics statement

Informed consent was obtained from all subjects involved in the study. No personalised data was collected during the study.

Disclosure statement

No potential conflict of interest is reported by the author(s).

About the authors

Daniel Hartmuth is teacher on secondment and PhD candidate at the Institute of Biology Education of the Faculty for Biology, LMU Munich.

Rifat Mehreen Amin is a PhD student working in Human-Computer Interaction. Her work focuses on advancing visualisation tools through the integration of Machine Learning and Artificial Intelligence. Collaborating at the intersection of Human-Computer Interaction (HCI) and AI, her projects often involve big data analytics, exploring innovative ways to improve user experiences and decision-making processes.

Daniel Schlichting is a Masters student and student assistant of StabLab (Statistical Consulting Unit) at LMU Munich.

Svenja Schött is a PhD student working in Human-Computer Interaction. Her work focuses on advancing visualisation tools through the integration of Machine Learning and Artificial Intelligence. Collaborating at the intersection of Human-Computer Interaction (HCI) and AI.

Jonathan Haudenschild is a Masters student working in Human-Computer Interaction.

Helmut Küchenhoff is a Professor and Head of StabLab (Statistical Consulting Unit) at LMU Munich.

Birgit J. Neuhaus is a Professor and chair holder of the Chair for Biology Education at LMU Munich, Germany. Her research interests are mainly connected to biology teachers' professional competence, instructional quality of biology lessons and simulations in biology teacher education.

ORCID

Daniel Hartmuth http://orcid.org/0009-0002-8063-307X

Data availability statement

Information on and queries about the data can be obtained from the authors of this article.

References

Abell, S. K. (2008). Twenty Years Later: Does pedagogical content knowledge remain a useful idea?. *International Journal of Science Education*, *30*(10), 1405–1416. https://doi.org/10.1080/09500690802187041

Alonzo, A. C., Berry, A., & Nilsson, P. (2019). Unpacking the complexity of science teachers' PCK in action: Enacted and personal PCK. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 273–288). Springer.

Alonzo, A. C., Kobarg, M., & Seidel, T. (2012). Pedagogical content knowledge as reflected in teacher–student interactions: Analysis of two video cases. *Journal of Research in Science Teaching*, *49*(10), 1211–1239. https://doi.org/10.1002/tea.21055

Altrichter, H., Posch, P., & Spann, H. (2018). *Lehrerinnen und Lehrer erforschen ihren Unterricht* [Teachers research their lessons] (5., grundlegend überarbeitete Auflage). Klinkhardt.

Bastian, J., Combe, A., & Langer, R. (2016). *Feedback-Methoden: Erprobte Konzepte, evaluierte Erfahrungen* [Feedback methods: Tried and tested concepts, evaluated experiences] (4., erweiterte und überarbeitete Auflage). *Pädagogik*. Beltz. http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-407-25754-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(), 1–48.

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

Behling, F., Förtsch, C., & Neuhaus, B. J. (2022). Using the plan–teach–reflect cycle of the refined consensus model of PCK to improve pre-service biology teachers' personal PCK as well as their motivational orientations. *Education Sciences*, *12*(10), 654. https://doi.org/10.3390/educsci12100654

Berry, A., Friedrichsen, P., & Loughran, J. (Eds.) (2015). *Teaching and learning in science series. Re-examining pedagogical content knowledge in science education*. Routledge. http://www.tandfebooks.com/isbn/9781315735665

Bijlsma, H., van der Lans, R., Mainhard, T., & den Brok, P. (2021). A reflection on student perceptions of teaching quality from three psychometric perspectives: CCT, IRT and GT. *Student Feedback on Teaching in Schools, 15*.

Bijlsma, H. J., Glas, C. A., & Visscher, A. J. (2022). Factors related to differences in digitally measured student perceptions of teaching quality. *School Effectiveness and School Improvement*, *33*(3), 360–380. https://doi.org/10.1080/09243453.2021.2023584

Bijlsma, H. J., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, *28*(2), 217–236. https://doi.org/10.1080/1475939X.2019.1572534

Blömeke, S., Hsieh, F. -J., Kaiser, G. , & Schmidt, W. H. (2014). *International perspectives on teacher knowledge, beliefs and opportunities to learn*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6437-8

Blömeke, S., & Kaiser, G. (2014). Theoretical framework, study design and main results of TEDS-M. In S. Blömeke, F.-J. Hsieh, G. Kaiser, & W. H. Schmidt (Eds.), *Advances in mathematics education. International perspectives on teacher knowledge, beliefs and opportunities to learn*. (pp. 19–47). Springer Netherlands. https://doi.org/10.1007/978-94-007-6437-8_2

Brühwiler, C., Helmke, A., & Schrader, F. -W. (2017). Determinanten der Schulleistung. [Determinants of school performance]. In M. K. Schweer (Ed.), *Lehrer-Schüler-Interaktion* (pp. 291–314). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15083-9_13

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Kam Ho Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., … Wilson, C. D. (2019). The refined consensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77–94). Springer Nature Singapore. https://doi.org/10.1007/978-981-13-5898-2_2

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

Chan, K. K. H., & Yung, B. H. W. (2018). Developing pedagogical content knowledge for teaching a new topic: More than teaching experience and subject matter knowledge. *Research in Science Education*, *48*(2), 233–265. https://doi.org/10.1007/s11165-016-9567-1

Davidowitz, B., & Potgieter, M. (2016). Use of the Rasch measurement model to explore the relationship between content knowledge and topic-specific pedagogical content knowledge for organic chemistry. *International Journal of Science Education*, *38*(9), 1483–1503. https://doi.org/10.1080/09500693.2016.1196843

de Kempenaer. (2017). *CMT level III 2018: The integration of technical anaylsis*. John Wiley.

DeWitt, P. (2018). Using the visible learning research to influence collaborative leadership. *Education Sciences*, *8*(4), 219. https://doi.org/10.3390/educsci8040219

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* [Research methods and evaluation in the social and human sciences]. Springer. https://doi.org/10.1007/978-3-642-41089-5

Douville, S., Grandjean Targos, P. T., Jones, N. D., Knight, C., & Azzam, T. (2025). Data visualization expert lessons learned: Implications for program evaluators. *American Journal of Evaluation*. Advance online publication. https://doi.org/10.1177/10982140241290744

Drits-Esser, D., Gess-Newsome, J., & Stark, L. A. (2017). Examining the sustainability of teacher learning following a year-long science professional development programme for inservice primary school teachers. *Professional Development in Education*, *43*(3), 375–396. https://doi.org/10.1080/19415257.2016.1179664

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, *23*(7), 5–12. https://doi.org/10.3102/0013189X023007005

Elstad, E., Lejonberg, E., & Christophersen, K. -A. (2017). Student evaluation of high-school teaching: Which factors are associated with teachers' perception of the usefulness of being evaluated? *Journal for Educational Research Online*, *9*, 99–117. https://doi.org/10.25656/01:12968

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications*. Springer.

Felser, G. (2015). *Werbe- und Konsumentenpsychologie [Advertising and consumer psychology]*. Springer. https://doi.org/10.1007/978-3-642-37645-0

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

Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, *42*, 91–99. https://doi.org/10.1016/j.stueduc.2014.04.003

Gaertner, H., & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(2), 159–182. https://doi.org/10.1007/s11092-018-9277-5

Gess-Newsome, J., & Lederman, N. G. (Eds.) (1999). *Science & technology education library*. Vol. *6*. *Examining pedagogical content knowledge: The construct and its implications for science education*. Kluwer Academic. http://www.loc.gov/catdir/enhancements/fy0813/99040998-d.html

Gess-Newsome, J., Taylor, J. A., Carlson, J., Gardner, A. L., Wilson, C. D., Stuhlsatz, M. A.. (2019). Teacher pedagogical-content knowledge, practice, and student achievement. *International Journal Ofscience Education*, *41*(7), 944–963. https://doi.org/10.1080/09500693.2016.1265158

Gould, S. J. (2007). *Punctuated equilibrium* (1. paperback ed.). Belknap Press of Harvard University Press.

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education. Professional development and practice series*. Teachers College Press.

Guo, Q., & Chen, Y. (2023). The effects of visual complexity and task difficulty on the comprehensive cognitive efficiency of cluster separation tasks. *Behavioral Sciences*, *13*(10), Article 827. https://doi.org/10.3390/bs13100827

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://www.taylorfrancis.com/books/9780203887332

Hattie, J., & Clarke, S. (2019). *Visible learning: Feedback*. Routledge. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5584342

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., & Lenske, G. (2013). Unterrichtsdiagnostik als Voraussetzung für Unterrichtsentwicklung [Lesson diagnostics as a prerequisite for lesson development.]. *Beiträge zur Lehrerbildung*, *31,* 214–233. https://doi.org/10.25656/01:13848

Hense, J. (Ed.). (2013). *Forschung über Evaluation: Bedingungen, Prozesse und Wirkungen*. [Research on evaluation: Conditions, processes and effects.] Waxmann. https://elibrary.utb.de/doi/book/10.31244/9783830978299

Irmer, M., Traub, D., Böhm, M., Förtsch, C., & Neuhaus, B. J. (2023). Using video-based simulations to foster pPCK/ePCK—New thoughts on the refined consensus model of PCK. *Education Sciences*, *13*(3), 261. https://doi.org/10.3390/educsci13030261

Irmer, M., Traub, D., & Neuhaus, B. J. (2024). Fostering ePCK in pre-service biology teacher education using adaptive scaffolding in a video-based simulation. *Cogent Education*, *11*(1), 2422272. https://doi.org/10.1080/2331186X.2024.2422272

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment* (1st ed.). Little Brown Spark.

Kämpfe, N. (2009). Schülerinnen und Schüler als Experten für Unterricht [Pupils as experts for lessons]. *Die deutsche Schule*, *101,* S.149–S.163. https://doi.org/10.25656/01:25587

Keller, J., Kwasnicka, D., Klaiber, P., Sichert, L., Lally, P., & Fleig, L. (2021). Habit formation following routine-based versus time-based cue planning: A randomized controlled trial. *British Journal of Health Psychology*, *26*(3), 807–824. https://doi.org/10.1111/bjhp.12504

Keller, M. M., Neumann, K., & Fischer, H. E. (2017). The impact of physics teachers' pedagogical content knowledge and motivation on students' achievement and interest. *Journal of Research in Science Teaching*, *54*(5), 586–614. https://doi.org/10.1002/tea.21378

Kirschner, S., Taylor, J., Rollnick, M., Borowski, A., & Mavhunga, E. (2015). Gathering evidence for the validity of PCK measures: Connecting ideas to analytic approaches. In *Re-examining pedagogical content knowledge in science education* (pp. 229–242). Routledge.

Kolb, A. Y., & Kolb, D. A. (2005). Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning & Education*, *4*(2), 193–212. https://doi.org/10.5465/amle.2005.17268566

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

Kubina, R. M., Kostewicz, D. E., Brennan, K. M., & King, S. A. (2017). A critical review of line graphs in behavior analytic journals. *Educational Psychology Review*, *29*(3), 583–598. https://doi.org/10.1007/s10648-015-9339-x

LeBreton, J. M., Moeller, A. N., & Wittmer, J. L. S. (2023). Data aggregation in multilevel research: Best practice recommendations and tools for moving forward. *Journal of Business and Psychology*, *38*(2), 239–258. https://doi.org/10.1007/s10869-022-09853-9

Lederman, N. G., & Gess-Newsome, J. (1999). Reconceptualizing secondary science teacher education. In J. Gess-Newsome & N. G. Lederman (Eds.), *Science & technology education library: Vol. 6. Examining pedagogical content knowledge: The construct and its implications for science education* (Vol. 6, pp. 199–213). Kluwer Academic. https://doi.org/10.1007/0-306-47217-1_8

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Science & technology education library: Vol. 6. Examining pedagogical content knowledge: The construct and its implications for science education* (Vol. 6, pp. 95–132). Kluwer Academic. https://doi.org/10.1007/0-306-47217-1_4

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

Marks, R. (1990). Pedagogical content knowledge: From a mathematical case to a modified conception. *Journal of Teacher Education*, *41*(3), 3–11. https://doi.org/10.1177/002248719004100302

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology*, *76*(5), 707–754. https://doi.org/10.1037/0022-0663.76.5.707

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal*, *21*(2), 341–366. https://doi.org/10.2307/1162448

McGuinness, B., & Wittgenstein, L. (Eds.). (2012). *Wittgenstein in Cambridge: Letters and documents 1911-1951* (4. ed. [Nachdr.]). Blackwell. https://doi.org/OeBC

Mientus, L., Hume, A., Wulff, P., Meiners, A., & Borowski, A. (2022). Modelling STEM teachers' pedagogical content knowledge in the framework of the refined consensus model: A systematic literature review. *Education Sciences*, *12*(6), 385. https://doi.org/10.3390/educsci12060385

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]*. Springer. https://doi.org/10.1007/978-3-642-20072-4

Morine-Dershimer, G., & Kent, T. (1999). The complex nature and sources of teachers' pedagogical knowledge. In J. Gess-Newsome & N. G. Lederman (Eds.), *Science & technology education library: Vol. 6. Examining pedagogical* *content knowledge: The construct and its implications for science education* (Vol. 6, pp. 21–50). Kluwer Academic. https://doi.org/10.1007/0-306-47217-1_2

Morris, T. H. (2020). Experiential learning – A systematic review and revision of Kolb's model. *Interactive Learning Environments*, *28*(8), 1064–1077. https://doi.org/10.1080/10494820.2019.1570279

Mueller, F. F., Kyburz, P., Williamson, J. R., Sas, C., Wilson, M. L., Dugas, P. T., & Shklovski, I. (Eds.). (2024). *Proceedings of the CHI conference on human factors in computing systems*. ACM.

Neuhaus, B. J. (2021). Unterrichtsqualität aus der Perspektive der Biologiedidaktik [Teaching quality from the perspective of biology didactics]. *Unterrichtswissenschaft*, *49*(2), 273–283. https://doi.org/10.1007/s42010-021-00114-x

Park, S. (2019). Reconciliation between the refined consensus model of PCK and extant PCK models for advancing PCK research in science. In A. Hume, R. Cooper, & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 119–130). Springer Nature Singapore. https://doi.org/10.1007/978-981-13-5898-2_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

Röhl, S. (2021). Effects of student feedback on teaching and classes: An overview and meta-analysis of intervention studies. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools* (pp. 139–156). Springer International. https://doi.org/10.1007/978-3-030-75150-0_9

Rollett, W., Bijlsma, H., & Röhl, S. (Eds.) (2021). *Student feedback on teaching in schools*. Springer International. https://doi.org/10.1007/978-3-030-75150-0

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *American Educational Research Journal*, *28*(1), 63–88. https://doi.org/10.3102/00028312028001063

Scheipl, F., Greven, S., & Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomialregression in additive and linear mixed models. *Computational Statistics &data Analysis*, *52*(7), 3283–3299. https://doi.org/10.1016/j.csda.2007.10.022

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014). Applying the User Experience Questionnaire (UEQ) in different evaluation scenarios. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, & A. Marcus (Eds.), *Lecture notes in computer science. Design, user experience, and usability. theories, methods, and tools for designing the user experience* (Vol. 8517, pp. 383–392). Springer International. https://doi.org/10.1007/978-3-319-07668-3_37

Schwabish, J. A. (2021). *Better data visualizations: A guide for scholars, researchers, and wonks*. Columbia University Press.

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 preservice teachers. *American Educational Research Journal*, *51*(4), 739–771. https://doi.org/10.3102/0002831214531321

Shavelson, R. J. (2012). Assessing business-planning competence using the collegiate learning assessment as a prototype. *Empirical Research in Vocational Education and Training*, *4*(1), 77–90. https://doi.org/10.1007/BF03546509

Shavelson, R. J. (2020). Research on teaching and the education of teachers: Brokering the gap. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *38*(1), 37–53. https://doi.org/10.25656/01:21773

Shavelson, R. J. (2023). Reflections on transfer of competence assessment to teaching and learning in higher education. *Zeitschrift Für Erziehungswissenschaft*, *26*(3), 585–599. https://doi.org/10.1007/s11618-023-01164-w

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Shulman, L. S. (1987). Knowledge and teaching:foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Skinner, B. F. (1959). *Cumulative record.* Appleton-Century-Crofts.

Smith, P. S., & Banilower, E. R. (2015). Assessing PCK: A new application of the uncertainty principle. In *Re-examining pedagogical content knowledge in science education* (pp. 88–103). Routledge.

Sorge, S., Kröger, J., Petersen, S., & Neumann, K. (2019). Structure and development of pre-service physics teachers' professional knowledge. *International Journal of Science Education*, *41*(7), 862–889. https://doi.org/10.1080/09500693.2017.1346326

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching. *Review of Educational Research*, *83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Suresh, H., Gomez, S. R., Nam, K. K., & Satyanarayan, A. (2021). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. Drucker (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445088

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder* (1st ed.). Random House.

Taleb, N. N. (2018). *Skin in the game: Hidden asymmetries in daily life* (International ed.). Random House.

Tobin, K., & McRobbie, C. (1999). Pedagogical content knowledge and co-participation in science classrooms. In J. Gess-Newsome & N. G. Lederman (Eds.), *Science & technology education library: Vol. 6. Examining pedagogical content knowledge: The construct and its implications for science education* (Vol. 6, pp. 215–234). Kluwer Academic. https://doi.org/10.1007/0-306-47217-1_9

van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, *30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110. https://doi.org/10.1016/j.chb.2018.07.027

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

Wisniewski, B., Röhl, S., & Fauth, B. (2022). The perception problem: A comparison of teachers' self-perceptions and students' perceptions of instructional quality. *Learning Environments Research*, *25*(3), 775–802. https://doi.org/10.1007/s10984-021-09397-4

Wisniewski, B., & Zierer, K. (2020). Empirische Arbeit: Entwicklung eines Online-Fragebogens zur Erhebung von Unterrichtsqualität durch Lernendenfeedback und erste Validierungsschritte. [Empirical work: Development of an online questionnaire to survey teaching quality through learner feedback and initial validation steps]. *Psychologie in Erziehung Und Unterricht*, *67*(2), 138–155. https://doi.org/10.2378/peu2020.art10d

Wisniewski, B., & Zierer, K. (2021). Functions and success conditions of student feedback in the development of teaching and teachers. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools* (pp. 125–138). Springer International. https://doi.org/10.1007/978-3-030-75150-0_8

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction*, *66*, 101303. https://doi.org/10.1016/j.learninstruc.2020.101303

Wisniewski, B., Zierer, K., & Hattie, J. (2019). 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