Social Psychology of Education (2020) 23:673–699
https://doi.org/10.1007/s11218-020-09554-2

Who sees the most? Differences in students' and educational research experts' first impressions of classroom instruction

Lukas Begrich[1] · Benjamin Fauth[2,3] · Mareike Kunter[1]

Received: 13 June 2019 / Accepted: 29 March 2020 / Published online: 18 April 2020

Abstract

In recent decades, the assessment of instructional quality has grown into a popular and well-funded arm of educational research. The present study contributes to this field by exploring first impressions of untrained raters as an innovative approach of assessment. We apply the thin slice procedure to obtain ratings of instructional quality along the dimensions of cognitive activation, classroom management, and constructive support based on only 30 s of classroom observations. Ratings were compared to the longitudinal data of students taught in the videos to investigate the connections between the brief glimpses into instructional quality and student learning. In addition, we included samples of raters with different backgrounds (university students, middle school students and educational research experts) to understand the differences in thin slice ratings with respect to their predictive power regarding student learning. Results suggest that each group provides reliable ratings, as measured by a high degree of agreement between raters, as well predictive ratings with respect to students' learning. Furthermore, we find experts' and middle school students' ratings of classroom management and constructive support, respectively, explain unique components of variance in student test scores. This incremental validity can be explained with the amount of implicit knowledge (experts) and an attunement to assess specific cues that is attributable to an emotional involvement (students).

Keywords Thin slices ratings · Instructional quality · Predictive validity · Incremental validity

This research was supported by Grants from the Federal Ministry of Education and Research (BMBF).

✉ Lukas Begrich
begrich@psych.uni-frankfurt.de

[1] Department of Psychology, Goethe-University Frankfurt, Frankfurt am Main, Germany

[2] Institute for Educational Analysis (IBBW), Stuttgart, Germany

[3] University of Tübingen, Tübingen, Germany

1 Introduction

The present study explores the potential of first impressions for assessing instructional quality. A research paradigm used in personality research, the thin slice procedure taps into first impressions by having people rate very short samples of the behavior of target persons (Ambady and Rosenthal 1992; Ambady et al. 2000). Applying this technique to classroom observations, prior studies have found evidence for high reliability as well as different indications of validity for thin slices ratings by untrained raters based on only a few seconds of observation (Ambady and Rosenthal 1993; Babad 2005; Begrich et al. 2017, 2019). Moreover, some initial evidence suggests that thin slice ratings of instructional quality can serve as significant predictors for student learning (Begrich et al. 2017). However, the underlying cognitive processes that produce these ratings are not yet well understood (Wood 2014). In particular, it is unclear whether thin slices ratings actually assess something specific about instructional quality or instead reflect global judgments related to personality or physical appearance of the teacher. In the present study, we vary rater samples in order to evaluate the specific predictive power that ratings from middle school students, university students and experts from educational research have for student learning. Intuitive judgements like thin slice ratings (Ambady 2010) are known to benefit both from domain-specific expertise (e.g. Dane and Pratt 2007; Dane et al. 2012) as well as special attunement to certain social stimuli because of their relevance for survival and well-being (Ambady et al. 2000). Therefore, if thin slices ratings reflect an assessment of information specific to and diagnostic for instructional quality, then we expect ratings of middle school students as well as educational research experts to have incremental validity in terms of predicting students' learning.

1.1 The thin slices procedure in the context of instructional research

The thin slices research paradigm was designed to investigate the accuracy of first impressions (Ambady and Rosenthal 1992; Ambady et al. 2000). Here, a “thin slice” is defined as an excerpt of expressive behavior that is sampled from any communication channel of the behavioral stream (Ambady et al. 2000). In most studies, thin slices are presented in the form of short (10-s to 5-min) video clips that provide visual, and sometimes auditory, information about a target person's expressive behavior. Typically, these short video clips are shown to samples of untrained observers, who judge the target person on the respective construct of interest solely based on their brief observations, i.e. their first impressions (Ambady et al. 2000). This procedure has been applied to various psychological constructs. Thin slice ratings have often been shown to be highly accurate in terms of agreement between the observers (i.e. reliability) and significant correlations with self-reports, reports of familiar persons or standardized tests (see e.g. Fowler et al. 2009; Friedman et al. 2007; Holleran et al. 2009; Oltmanns et al. 2004; Tskhay et al. 2017). For example, Borkenau et al. (2004) used 4-min (average length) video clips showing the target person engaged in different activities (e.g. introducing themselves) as the basis for thin slice ratings of the Big Five personality traits and intelligence. They found significant correlations between thin slices ratings of personality and reports of familiar persons as well as between thin slices ratings of intelligence and scores from two standardized intelligence tests (r = 0.41 and 0.53). Similarly, Carney et al. (2007) found significant correlations between thin slices ratings of negative affect, extraversion, conscientiousness, and intelligence that were based on 5-s observations as well as targets' self-reports and test scores. Moreover, studies applying the thin slices procedure have repeatedly shown that first impressions are a sufficient basis for detecting signs of personality or developmental disorders in strangers (Fowler et al. 2009; Manson et al. 2018; Oltmanns and Turkheimer 2006; Walton and Ingersoll 2016). In more applied domains, thin slices ratings have frequently shown predictive validity regarding different practical outcomes. In organizational psychology, for example, Visser and Matthews (2005) found thin slices ratings based on 30-s observations of sales pitches could significantly predict customer satisfaction and supervisor assessments. Interestingly, thin slice ratings were also found to accurately assess the quality of interactions between two or more people. Carrère and Gottman (1999) found that thin slice ratings of 3-min video clips of couples having a discussion from 6 years earlier could predict divorce rates. Similarly, in a study by Lambert et al. (2014) raters could accurately identify which partner in a couple cheated on the other by watching 3- to 5-min clips of the couple's interactions. Jung (2016) also provides evidence that a team's work performance could be predicted by thin slice ratings of the quality of interactions between different team members based on 15 min of observation.

Few studies into education have used the thin slices procedure to evaluate classroom instruction. These studies also greatly differed both in their aim as well as instruction-related construct under investigation (Ambady and Grey 2002; Ambady and Rosenthal 1993; Praetorius et al. 2015; Pretsch et al. 2013; Strong et al. 2011). In their pioneer study, Ambady and Rosenthal (1993) investigated the effect of observation duration by presenting undergraduates with 6-, 15- or 30-s muted video clips of teachers at work. The participants were asked to judge the teachers regarding their physical attractiveness and personality traits such as "accepting", "likable" or "supportive" as well as attributes related to their profession such as "competent" or "professional". The study found both strong agreement among the raters and significant correlations between the ratings and evaluations of teachers' effectiveness given by their students and supervisors. Interestingly, regardless of video length, there were no significant differences between the groups in terms of agreement or correlations with the teachers' effectiveness as judged by their students and supervisors. Mainhard et al. (2014) compared personality ratings of teachers provided by their students and another teacher's students based on 5 min of observation. They found substantial overlap between these ratings, concluding that students' first impressions are quite accurate regarding a teacher's personality. Pretsch et al. (2013) investigated whether first impressions could be used to predict teachers' well-being and job satisfaction. They found significant correlations between thin slices ratings based on 60-s muted videos showing teachers in action and self-reported well-being and job satisfaction 3 and 6 years later. Earlier findings by Babad et al. (1991) suggest that first impressions of interactions between teachers and students are an accurate measure for the former's expectations of the latter's achievement. Moreover, Babad (2005) found evidence that it is even possible to assess teachers' instructional practices via thin slice ratings under certain circumstances. Based on ten-s video clips of teachers talking to the entire class, ratings of teachers' differential treatment of high- versus low-achievers could predict ratings of this tendency given by the students of these classes. This was only the case if the raters of the thin slices were students, too, however.

1.2 Thin slices as intuitive judgments

Dual-process theories of social cognition offer an explanation of the intriguing high accuracy of thin slices ratings (e.g. Brewer 1988; Kahneman 2011; Stanovich and West 2000). Applied to perception of people, these theories postulate two different stages of information processing that differ in their underlying cognitive processes (Ambady 2010; Kunda and Thagard 1996). Cognitive processes underlying thin slices ratings refer to the first stage of person perception that is characterized by automaticity and occurrence of outside awareness. In contrast, the second stage of perceiving a person builds on richer information about others and is characterized by deliberate und conscious cognitive processing. Impressions are effortlessly formed in the first stage and form the basis of judgments regarding others' personality, affect, and behavioral tendencies. Interpersonal relations are also driven by non-verbal cues that are decoded from others' expressive behavior (Ambady 2010; Feldman 1981; Fiske and Neuberg 1990; Gingerich et al. 2011). Ambady (2010) pointed out that thin slices ratings that take advantage of first-stage processes have all the characteristics of intuitive judgments. First, in contrast to cognitive processes on the second stage, thin slice ratings are efficient; they are unaffected by distractions and open to parallel processing (Ambady 2010; Costanzo and Archer 1989; Patterson and Stockbridge 1998). Second, the accuracy of thin slices ratings seems to suffer from deliberation, which is attention to the intuitive information processing through verbalization, for example (Ambady 2010; Murphy and Balzer 1986). A possible explanation for these findings is that deliberation leads to an idiosyncratic emphasis regarding the importance of certain cues (Ambady 2010). Intuitive judgments also rely on implicit and tacit, rather than declarative, knowledge. Therefore, intuitive judgments become more accurate if they are made within a familiar domain about which a social perceiver has appropriate implicit knowledge (Ambady 2010). Recent conceptualizations of expertise emphasize its domain-specificity as well as its dependency on tacit knowledge. Therefore, an expert who is classically defined as someone "whose level of performance exceeds that of most others" (Cianciolo et al. 2006, p. 614) is now often seen as someone who knows more than she or he can tell (Cianciolo et al. 2006). Consequently, growing expertise is assumed to be reflected in an increasing reliance on intuition (Dreyfus and Dreyfus 2005). Actually, domain experts seem to especially profit from intuitive judgments and decision-making (e.g. Dane and Pratt 2007; Dane et al. 2012). Moreover, this benefit increases with years of intense activity and practice in a certain domain (Dreyfus and Dreyfus 2005; Ericsson and Charness 1994; Simon 1987). Empirical findings documenting a benefit from domain expertise regarding the accuracy of thin slices ratings are presented in a study by Correll et al. (2007). They show that police officers are more accurate than other community members in decisions about whether to shoot or not based on their first impressions. Hence, if one wanted to use thin slice ratings to assess instructional quality, educational experts, i.e., scholars from research on instructional quality, would likely provide particularly accurate judgments. Finally, intuitive judgments seem to benefit if they are affectively connected to a social perceiver (Ambady 2010). Gibson's "Event Action Approach" (Gibson 1979; cited from Ambady et al. 2000) offers an explanation for this phenomenon. The theory states that people must have a certain *attunement* to social stimuli (*affordances*) to take notice and intuitively process it. Therefore, the degree of attunement to certain social stimuli depends on either prior experiences or a special need of the social perceiver to assess these stimuli because of their relevance for survival or well-being (Ambady et al. 2000). Thus, we hypothesize that recruiting students is especially suitable for assessing instructional quality via thin slices ratings. Due to their attunement to instructional features connected to their well-being, we expect that middle school students are more sensitive and better able to detect cues indicating how supportive a teacher is. Babad's (2005) study supports this hypothesis. The finding that only students are able to infer the overall teachers' differential treatment of high- versus low-achievers from ten-s video clips of teachers talking to the whole class could be interpreted as a sign of their special attunement to cues indicating this tendency in teachers. In contrast, adults whose well-being does not depend on such teacher behaviors seem less capable of giving valid ratings based on their first impressions. Taken together, one can expect intuitive judgments to be more accurate for social perceivers that either are familiar with and have extensive tacit knowledge in a certain domain or those for which detecting social stimuli in a given domain has an affective component.

1.3 Instructional quality

Educational research has identified the quality of instruction as a key factor in students' learning progress (e.g. Hattie 2009; Pianta and Hamre 2009). In fact, the theoretical conceptualization and assessment of instructional quality has become a frequently addressed and highly funded direction of research (e.g. Hattie 2009; Kane et al. 2012; Rivkin et al. 2000; Seidel and Shavelson 2007). One approach for assessing the quality of instruction is to use ratings by external observers, which are often based on classroom videos and have been repeatedly shown to have predictive value regarding student learning, at least when scores of different observers and/or observations are averaged (Kane et al. 2012; Praetorius et al. 2012, 2014). Over the past several decades, various instruments for observational measures have been developed, which strongly differ in their underlying theories of instructional quality, the scope of teaching aspects assessed, the dimensionality of measured constructs, and subject-relatedness (e.g. Brophy and Good 1986; Gargani and Strong 2014; Pianta and Hamre 2009). What most of these approaches have in common is that judgments require an inference from single observations to outlasting instructional quality features (Kane et al. 2012; Klieme et al. 2009). Raters must be extensively trained to obtain reliable and valid scores (Kane et al. 2012), making many observational rating procedures time-consuming and expensive (Gargani and Strong 2014).

The Three Basic Dimensions are a well-established model of instructional quality in German-speaking countries (Klieme 2006; Kunter and Baumert 2006; Praetorius et al. 2018). The model is both empirically and theoretically founded on three central dimensions of instructional quality: cognitive activation, classroom management, and constructive support. These dimensions are generic as to be applicable across school subjects (Praetorius et al. 2018) and refer to aspects of the learning-related interactions between teachers and students, each consisting of a set of sub-dimensions describing more specific instructional practices (Kunter and Baumert 2006). Cognitive activation refers to the potential of a teacher's instruction to stimulate insightful and deep engagement with learning materials by building on and challenging students' existing concepts and knowledge structures (e.g. Baumert et al. 2010; Lipowsky et al. 2009). It can be achieved by asking students stimulating questions and was found, for example, to be an important mediator when explaining students' learning based on teachers' pedagogical content knowledge (Baumert et al. 2010). Successful classroom management results in an efficient use of time by both structuring learning time as well as preventing disruptions. One instructional practice important for successful classroom management is to establish clear rules and procedures (Emmer and Evertson 2013; Emmer and Stough 2001; Marzano et al. 2003). An effective use of learning time is empirically associated with students' achievements and motivation (Seidel and Shavelson 2007; Rakoczy et al. 2007). Finally, constructive support refers to a positive, student-oriented style of interaction created by the teacher that fosters positive relationships between a teacher and their students (Kunter and Baumert 2006). An important practical approach to providing support in the classroom is to give constructive and appreciative feedback to students (Klieme et al. 2009). Studies show that supportive instruction is positively related to emotional and motivational student outcomes and that it is an important factor in students' well-being and subject-related interest (Fauth et al. 2014b; Kunter et al. 2013; Praetorius et al. 2018). Numerous studies have investigated the three basic dimensions with different approaches of assessment (teacher ratings, student ratings, ratings by external observers), school forms, and subjects (Praetorius et al. 2018). Although the expected predictive effects are not consistently found across studies, the framework presents a comprehensive and parsimonious taxonomy of instructional quality relevant to different student outcomes by empirically connecting the identified dimensions to existing theories from educational science and psychology. The basic dimensions of instructional quality are purposefully generic to be relevant across subjects, school years and even school forms (Praetorius et al. 2018). When it comes to cognitive activation of instruction, however, there is evidence that it cannot be investigated independently from the specific learning content (Praetorius et al. 2018) and that it may require more or longer observations for assessments to be reliable (Praetorius et al. 2014).

1.4 Present research

Considering the few promising results for the potential of thin slice ratings to assess instructional practices, we evaluate this method as an economical observational approach for assessing instructional quality along the three basic dimensions. In prior studies, we presented undergraduate psychology and educational science students with 30-s slices randomly drawn from classroom videos stemming from different video studies (Begrich et al. 2017, 2019). Untrained participants rated these observations along the three basic dimensions. The only instruction they were given was to rely on their first impressions. All of the raters demonstrated a high degree of agreement in their ratings in every study. In a study applying this procedure to elementary school science classes, we found hints of convergent validity in terms of a clear pattern of overlap with ratings of trained observers that were based on the full 90-min classroom videos (Begrich et al. 2017). Moreover, thin slices ratings of classroom management and cognitive activation were predictive of students' learning. In two other studies (Begrich et al. 2019) that applied the same procedure to 8th grade math and 9th grade English classes, confirmatory factor analysis revealed that thin slices ratings could differentiate between the three dimensions of instructional quality.

However, we found high correlations between the thin slices ratings of instructional quality and physical attractiveness and various personality traits (Begrich et al. 2017). In addition, inter-factor correlations for thin slices ratings were higher than correlations between the dimensions of instructional quality found in other studies working with student ratings, teacher self-reports, and task analyses (e.g., Fauth et al. 2014b; Künsting et al. 2016; Kunter and Voss 2011). Therefore, the question remains if thin slices ratings assess something specific to instructional quality or if they are instead driven by more global, personality-related, and perhaps superficial judgments. Moreover, there are good reasons to assume that trained observers' ratings based on longer observations could also be partly affected by judgmental biases (Praetorius et al. 2012). For instance, the classic study by Naftulin et al. (1973) demonstrated how the presenter's persona can affect the reception of specialized lectures (known as "Dr. Fox effects"). Halo effects, the tendency to overestimate connections between different personality traits, have also been found to affect rating data (Feeley 2002). Therefore, correlations between thin slice ratings and ratings from trained raters based on longer observations may be spurious due to variables not assessed in the studies. These variables could be impressions related to a teachers' personality, e.g. "charisma".

In this study, we address these issues by varying rater samples to investigate if thin slices ratings of instructional quality benefit from domain-specific expertise and attunement. This approach builds on findings that suggest that domain-specific implicit knowledge allows more accurate intuitive judgments like thin slices ratings, for example, in the respective domain (see e.g. Ambady 2010; Correll et al. 2007; Dane and Pratt 2007; Dane et al. 2012). In a similar manner, intuitive judgments seem to profit if they have an affective component on the side of the social perceiver, i.e. attunement (see Ambady et al. 2000; Babad 2005). Therefore, if first impressions of instruction contain information specific to and diagnostic for instructional quality features, we expect higher accuracy of thin slices ratings given by students (because of an attunement to features of their instruction relevant to their well-being) and experts in research on instruction (because of great implicit knowledge of instructional processes) compared to thin slices ratings of adults with neither special attunement nor expertise related to instruction (e.g. the adult psychology students typically used in thin slices studies). If thin slices ratings of instructional quality are distorted by results of the halo effect insofar as they mainly reflect judgments of personality and/or physical appearance, these judgments should not benefit from a student or expert perspective.

We investigated the following research questions:

1. Are the thin slices ratings of instructional quality given by university students, middle school students, and experts reliable (strong agreement within each rater group)? Because of the very strong agreement between adult raters found in previous studies, we expect reliability of thin slices ratings to be comparably high in all samples with Intraclass-Correlation Coefficient (ICC) scores higher than .70.
2. Are the thin slices ratings of instructional quality given by university students, middle school students, and experts valid for predicting students' learning? As our previous findings indicate the predictive validity of thin slices ratings of cognitive activation and classroom management, we expect the thin slices ratings of these dimensions to be predictive regarding student learning for all rater samples.
3. Is there an incremental validity of thin slices ratings given by middle school students and experts that is evident in unique contributions to the explanation of variance in students' test scores?

    a. Do thin slices ratings of instructional quality benefit from a special attunement to its cues? As instructional support is relevant to students' affect (Praetorius et al. 2018), we expect that middle school students are more sensitive and more likely to detect cues related to teacher support (Ambady et al. 2000; Babad 2005). Therefore, if thin slices ratings truly assess this dimension, middle school student ratings should better explain a unique component of variance in students' learning as compared to the other groups' ratings.
    b. Do thin slices ratings of instructional quality benefit from domain-specific expertise? As cognitive activation seems to be more difficult to assess based on single observations (Praetorius et al. 2014), we expect a benefit from domain expertise to be visible in thin slices ratings of this dimension. If thin slices ratings truly assess cognitive activation, we expect expert ratings to better explain a unique component of variance in students' learning as compared to the other groups.
    c. As findings from other studies indicate that classroom management can be assessed from the students' perspective as well as from trained observers (Clausen 2002; Fauth et al. 2014b; Waldis et al. 2010), we do not expect thin slices ratings of classroom management to benefit from domain-expertise or attunement. Therefore, we do not expect student or expert ratings of this dimension to explain unique variance in students' learning.

2 Method

2.1 Data basis

The data used in this study are classroom videos and student data from the German IGEL study (Hardy et al. 2011), which investigated science instruction in elementary school and tested the effects of two standardized instructional units about the topic "floating and sinking" on student learning (Decristan et al. 2015). We used classroom videos and student data of 23 classes from German public primary schools with a total of 431 students. The average age of student in this study was 8.8 years (SD = 0.50) with 49% female students. Science teachers in this study were, on average, 42.8 years old (SD = 9.2), with an average professional experience of 16.4 years (SD = 8.6). Teachers and students voluntarily participated in the study. The data collection was approved by the Ethics Committee of the Faculty of Psychology and Sports Sciences at a German University.

2.2 Stimulus material

Every lesson was split into three parts of equal length to create the stimulus material. Ten-s slices were randomly drawn from each one-third of each of the 23 classroom videos. Only video slices in which the teacher was clearly visible were used in this study. If the teacher was not visible in a selected clip, subsequent slices from the same one-third of a lesson were selected until the criterion was met. The stimulus material consisted of 30-s video clips showing three, 10-s episodes of each teacher. As the classroom videos show the standardized instructional units mentioned above, these episodes were comparable between the different classes. In most cases, raters saw the teachers either discussing the concepts of floating and sinking with their students or walking around supporting the students as they completed group work.

2.3 Rater samples

2.3.1 University students

A sample of nine (six female) undergraduate psychology students participated in the study in exchange for course credit. Participants self-reported no specific expertise in teaching or sound knowledge about research on instruction. The mean age of participants were 23.7, ranging between 20 and 29. Their ratings were also published in a previous paper (Begrich et al. 2017).

2.3.2 Middle school students

The student sample consisted of one class of 8th graders recruited at a German secondary school. 23 students (15 females) voluntarily participated in the study.

Students ranged in age between 13 and 15 years old. To equal the number of participants in each rater sample we randomly drew nine students (five females and 4 males) and included their data in the analyses.

2.3.3 Experts

An ad-hoc sample of educational research experts was recruited in a special interest group focused on topics related to instruction and professionalization at a German University. Of the nine experts participating in the study, seven were female. The average age of the expert group was 39.2, ranging between 27 and 63 years old. The sample consisted of three professors, with degrees in psychology, a combination of mathematics, psychology and educational science, and physics. Three participants in the sample hold doctoral degrees—two in psychology, and one in educational science. The remaining participants were PhD students with master's degrees in psychology, educational science or teacher training programs. Every participant self-reported dealing with issues of instructional quality in their research. On average, participants reported having professionally dealt with instructional quality for 12.2 years, ranging from 3 to 39 years. Participants rated their own expertise on instructional quality on a six-point Likert scale reporting an average score of 4.9.

2.4 Instruments

Students were given a standardized achievement test before and after the instructional units to assess their conceptual understanding of the material (adapted from Hardy et al. 2006, 2010). The post-test version of this achievement test consisted of 13 items (EAP/PV-reliability = 0.76). Besides the pre-test version (EAP/PV-reliability = 0.52) the following covariates were included to account for pre-existing differences between students: 1. Intelligence test (CFT, 20-R; Weiß 2006; 56 Items, Cronbachs $\alpha = 0.72$). 2. Standardized test of general scientific literacy (inspired by the TIMSS 2007 science test by Bos et al. 2007; 12 items, EAP/PV-reliability = 0.70).

All participants viewed the stimulus material and rated the quality of instruction based on six items. We adapted these items from existing instruments and used a six-point Likert scale to obtain the ratings. We measured cognitive activation, the teachers' ability to build on students' prior knowledge and to stimulate insightful learning processes, by using two items that asked for raters' impressions of teachers' success in leading students to an active and deep engagement with learning materials and content (see Table 1 for item wording). We also used two items to measure classroom management, the teachers' ability to efficiently use learning time by structuring instruction and preventing interruptions. These items ask raters for their impression regarding the frequency of interruption as well as teachers' ability to involve all students in the learning process. Finally, we measured constructive support provided by teachers using two items that asked raters for their impressions of the quality of teacher–student interactions as well as teachers' tendencies to help their students with understanding problems (see Table 1). We calculated scores for every scale by averaging raters' responses on each of the two items.

Table 1 Texts, reliabilities, means and standard deviations of items and scales for thin-slices ratings by adults, students, and experts

Item/scale	University students (n = 9)			Middle school students (n = 9)			Experts (n = 9)		
	ICC (2)	M	SD	ICC (2)	M	SD	ICC (2)	M	SD
Cognitive activation scale	.99*	4.02	0.90	.95*	3.40	0.54	.98*	3.65	0.73
The teacher succeeds in stimulating the students to an active engagement with the learning material	.90*	4.14	0.94	.80*	3.42	0.53	.86*	3.75	0.79
The teacher is leading its students to deeply engage with the learning content	.89*	3.89	0.87	.79*	3.38	0.57	.79*	3.49	0.73
Classroom management scale	.86*	3.95	0.92	.90*	3.55	0.58	.89*	3.76	0.76
Interruptions occur rarely	.89*	3.97	1.02	.83*	3.55	0.60	.80*	3.94	0.85
The teacher manages to involve all students in the learning process	.89*	3.93	0.95	.79*	3.55	0.63	.79*	3.53	0.78
Constructive support scale	.93*	4.16	0.95	.78*	3.67	0.61	.96*	3.51	0.69
The teacher helps his or her students if understanding problems occur	.89*	4.20	0.97	.81*	3.64	0.60	.67*	4.03	0.69
Interactions between teacher and students are characterized by appreciation and respect	.88*	4.16	1.00	.77*	3.71	0.67	.79*	3.82	0.76

*$p < .05$

Cognitive activation scale: McDonald's ω = .99 (adults); McDonald's ω = .95 (students); McDonald's ω = .98 (experts)

Classroom management scale: McDonald's ω = .86 (adults); McDonald's ω = .91 (students); McDonald's ω = .90 (experts)

Constructive support scale: McDonald's ω = .93 (adults); McDonald's ω = .8 (students); McDonald's ω = .94 (experts)

2.5 Procedures

University students, middle school students, and experts were separately tested within a group setting. While university students and experts participated in the study in rooms at a German university, middle school students were tested in their classroom. At the beginning of the session, participants were briefly instructed to only rely on their first impressions to give their ratings. Afterwards, participants had the opportunity to read the text of each item and to ask questions if there were any comprehension problems. The video clips were projected on a large screen and participants rated the quality of each teacher's instruction directly after watching each 30-s video clip.

3 Results

3.1 Reliability (research question 1)

We calculated average Intraclass-Correlation Coefficient (ICC) scores to check the reliability of the thin slices ratings of instructional quality given by each rater group. ICC (2; Lüdtke et al. 2009) scores quantify the agreement between different raters by relating the variance of rater's judgements regarding the same measured object to the variance calculated throughout all measurement objects. The lower the variance of judgements regarding single measurements is compared to the total variance, the closer the ICC is to 1 (Rosenthal 1987). Table 1 displays ICC scores, means and standard deviations for every item and sample as well as for the formed scales. On the item level, the ICC scores of the different rater samples are all comparably high, with slightly lower ICC scores for thin slices ratings in the expert group. This indicates a strong overall agreement between the raters within each group. On the scale level, however, ICC scores for classroom management and cognitive activation are slightly higher within the expert group compared to ICC scores within the adult and the student group. Together with the high omega coefficients, which indicate high internal consistency of the scales (McDonald 1970, 1999), these results show that the thin slices ratings are highly reliable. Summarily, the first impressions that raters derived from their 30-s observations are quite similar within each sample.

3.2 Preliminary analyses

The empirical means of thin slices ratings were slightly higher compared to the theoretical means of the six-point rating scales in each rater sample (Table 1). The standard deviations are consistently highest for university students' ratings and lowest for middle school students' ratings. Table 2 displays correlations between the dimensions of instructional quality within and between the rater samples. Within each sample, we found significant correlations between thin slices ratings of cognitive activation, classroom management and constructive support. Descriptively,

Table 2 Multitrait–multimethod matrix (bivariate correlations for three dimensions of instructional quality measured from three perspectives)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
University students									
(1) Cognitive activation									
(2) Classroom management	.83*								
(3) Constructive support	.83*	.72*							
Middle school students									
(4) Cognitive activation	.52*	.43*	.50*						
(5) Classroom management	.52*	.50*	.40	.84*					
(6) Constructive support	.48*	.46*	.41	.87*	.74*				
Experts									
(7) Cognitive activation	.52*	.58*	.50*	.48*	.27	.63*			
(8) Classroom management	.50*	.71*	.50*	.42*	.32	.58*	.81*		
(9) Constructive support	.61*	.53*	.80*	.32	.13	.36	.71*	.54*	

*$p < .05$

correlations between the three dimensions tend to be lower within the expert sample as compared to the university and middle school students. Experts seem to differentiate more between the different dimensions when rating instructional quality based on their first impressions than university and middle school students.

Taking the inter-correlations between the ratings given by the different rater samples into account, there seems to be a lack of discriminant validity in thin slices ratings of instructional quality (Campbell and Fiske 1959). It is unclear, however, if these high inter-correlations are only due to undifferentiated judgments of the observers. They might also be due to an actual high covariation of these quality dimensions in practice with teachers being more or less simultaneously successful in cognitively activating students and providing a structured classroom management as well as constructive support (Holzberger et al. 2019). To put the high inter-correlations of thin slices ratings in perspective, Table 3 displays the inter-correlations between ratings of the three quality dimensions given by trained observers based on the full classroom videos thin slices were sampled from in the present study. These ratings were obtained in the IGEL-Study (Hardy et al. 2011; see above; Fauth et al. 2014a). Obviously, as ratings of the three quality dimensions share less variance compared to thin slices ratings, trained observers who have seen the whole lessons

Table 3 Intercorrelations between three dimensions of instructional quality measured with ratings by trained observers based on observations of full classroom videos

	(1)	(2)	(3)
(1) Cognitive activation			
(2) Classroom management	.59*		
(3) Constructive support	.28	.67*	

*$p < .05$

seem to judge instruction more differentiated than observers that rate instructional quality based on their first impressions. In another study, however, using multilevel confirmatory factor analysis we actually could demonstrate that there is some differentiation of the basic dimensions in thin slices ratings (Begrich et al. 2019).

3.3 Predictive validity (research question 2)

Multilevel regression analyses were conducted using Mplus 7.4 to examine the predictive validity of thin slice ratings of instructional quality regarding students' learning (Muthén and Muthén 1998–2013). Students' scores on the post-test served as the measure of their learning. Thin slices ratings of cognitive activation, classroom management and constructive support were introduced in three separate models for each perspective as predictors at the classroom level (level 2). The covariates of students' learning (pre-test scores, intelligence, and scientific competence) were introduced in each model on the individual level (level 1; see Tables 3, 4, 5) as grand-mean centered predictors.

On the individual student-level, students with greater prior domain-specific knowledge, generally higher scientific literacy and higher intelligence tended to score higher on the achievement test after the standardized instructional unit (see Table 4). Due to the high correlations between the dimensions within each sample (see Table 2), we used university students' ratings of the three dimensions of instructional quality in three different models on the classroom level to understand whether differences in instructional quality between classes as assessed by thin slices ratings explained the differences in students' learning (see Table 4, models 1a–c). After controlling for covariates, we found that university students' thin slices ratings of classroom management and cognitive activation significantly predict students' learning. This is not the case for ratings of constructive support. Therefore,

Table 4 Multilevel regression analyses predicting students' knowledge on floating and sinking by thin-slices ratings of the three dimensions of instructional quality given by university students

Predictor	Thin-slices ratings university students		
	Model 1a	Model 1b	Model 1c
Individual level			
Pre-test	.24* (.05)	.24* (.05)	.24* (.05)
Intelligence (CFT)	.22* (.05)	.21* (.05)	.21* (.05)
Scientific competence	.23* (.04)	.23* (.04)	.23* (.04)
Classroom level			
Cognitive activation	.37* (.21)	–	–
$R^2_{between}$	.14 (.16)	–	–
Classroom management	–	.49* (.20)	–
$R^2_{between}$	–	.24 (.20)	–
Constructive support	–	–	.32 (.23)
$R^2_{between}$	–	–	.10 (.15)

Standardized regression weights; standard errors are in parentheses

*$p < .05$; one-tailed test

Table 5 Multilevel regression analyses predicting students' knowledge on floating and sinking by thin-slices ratings of the three dimensions of instructional quality given by middle school students

Predictor	Thin-slices ratings students		
	Model 2a	Model 2b	Model 2c
Individual level			
Pre-test	.24* (.05)	.24* (.05)	.24* (.05)
Intelligence (CFT)	.21* (.05)	.21* (.05)	.21* (.05)
Scientific competence	.23* (.04)	.23* (.04)	.23* (.04)
Classroom level			
Cognitive activation	.40* (.15)	–	–
R^2_{between}	.16 (.12)	–	–
Classroom management	–	.39* (.17)	–
R^2_{between}	–	.15 (.13)	–
Constructive support	–	–	.49* (.14)
R^2_{between}	–	–	.24 (.14)

Standardized regression weights; standard errors are in parentheses

*$p < .05$; one-tailed test

it seems that ratings of these dimensions by untrained university students based on 30 s of observation are sufficient to predict how much students learn in different classes. Descriptively, ratings of classroom management explain more variance in students' test scores than ratings of cognitive activation (24% versus 14%).

Our study assumed that the middle school students are particularly apt at rating teacher support due to their higher attunement. Table 4 lists the results for the models, in which middle school students' thin slices ratings of instructional quality function as level-2 predictors for student learning. Their thin slices ratings of every dimension of instructional quality are predictive with respect to student learning (see Table 5). Ratings of constructive support, however, explain the greatest component of variance in students' test scores (24%), while cognitive activation (16%) and classroom management (15%) are less explanatory for variance in students' test scores.

Table 6 shows the results of experts' thin slices ratings. Every dimension of instructional quality as rated by experts is significantly predictive of students' learning. Ratings of cognitive activation explain 12% of variance in students' test scores, while ratings of classroom management explain 42% and ratings of constructive support explain 13% of variance.

Overall, we thus find that ratings of instructional quality based on 30 s of observation are predictive of students' learning.

3.4 Incremental validity (research question 3)

As thin slices of every dimension of instructional quality given by all three samples explain significant components of variance in students' test scores with only one exception, it is important to examine if one perspective (experts, university or middle school students) explains a unique component of variance not accounted for

Table 6 Multilevel regression analyses predicting students' knowledge on floating and sinking by thin-slices ratings of the three dimensions of instructional quality given by experts

Predictor	Thin-slices ratings experts		
	Model 3a	Model 3b	Model 3c
Individual level			
Pre-test	.24* (.05)	.24* (.05)	.24* (.05)
Intelligence (CFT)	.22* (.05)	.21* (.05)	.22* (.05)
Scientific competence	.23* (.04)	.23* (.04)	.23* (.04)
Classroom level			
Cognitive activation	.33* (.13)	–	–
$R^2_{between}$	.12 (.08)	–	–
Classroom management	–	.65* (.15)	–
$R^2_{between}$	–	.42 (.22)	–
Constructive support	–	–	.36* (.20)
$R^2_{between}$	–	–	.13 (.15)

Standardized regression weights; standard errors are in parentheses

*$p < .05$; one-tailed test

by the other perspectives. Therefore, we tested three additional models, in which we simultaneously introduced thin slices ratings of all three rater samples for each dimension (see Table 7). When it comes to thin slices ratings of cognitive activation, no perspective explains a significant unique component of variance more than the other two. All three perspectives account for 21% in students' test scores. For classroom management, ratings given by experts explain a significant and additional component of variance in students' test scores. The three perspectives together explain 45% of variance here, which means that ratings by university students and middle school students explain an additional 3% of variance in student test scores. Similarly, when introducing thin slices ratings of constructive support given by all three perspectives, ratings by middle school students explain a significant and additional component of variance in students' test scores. Together, the three perspectives explain 29% of variance here. Thus, ratings by university students and experts account for additional 5% of variance in students' ratings.

4 Discussion

The present study explored the thin slices procedure as a potentially economical approach for assessing instructional quality. In prior studies, we found ratings of instructional quality given by untrained observers based on their first impressions to be highly reliable as well as valid under certain circumstances (Begrich et al. 2017, 2019). In this study, we investigated if thin slices ratings given by observers with a domain-specific expertise and/or a special attunement to cues signaling aspects of instructional quality show incremental validity in terms of explaining unique components of variance in students' learning. First, thin slices ratings were reliable in all three rater samples. Second, thin slices ratings of instructional quality appear to

Table 7 Multilevel regression analyses predicting students' knowledge on floating and sinking by thin-slices ratings of the three dimensions of instructional quality given by university students, middle school students and experts

Predictor	Thin-slices ratings: adults, students, experts		
	Model 4a	Model 4b	Model 4c
Individual level			
Pre-test	.24* (.05)	.24* (.05)	.24* (.05)
Intelligence (CFT)	.21* (.05)	.21* (.05)	.21* (.05)
Scientific competence	.23* (.04)	.24* (.04)	.23* (.04)
Classroom level			
Cognitive activation (university students)	.18 (.32)	–	–
Cognitive activation (middle school students)	.25 (.21)	–	–
Cognitive activation (experts)	.12 (.20)	–	–
$R^2_{between}$	.21 (.14)	–	–
Classroom management (university students)	–	.004 (.31)	–
Classroom management (middle school students)	–	.20 (.19)	–
Classroom management (experts)	–	.58* (.20)	–
$R^2_{between}$	–	.45 (.20)	–
Constructive support (university students)	–	–	–.01 (.31)
Constructive support (middle school students)	–	–	.43* (.17)
Constructive support (experts)	–	–	.26 (.35)
$R^2_{between}$	–	–	.29 (.17)

Standardized regression weights; standard errors are in parentheses

*$p < .05$; one-tailed test

provide valid predictions of student learning. Third, we found differences between the rater samples with respect to unique components of variance explained in students' learning. Raters' expertise and attunement can explain these differences. We view this last finding as an indication that thin slices ratings of instructional quality reflect more than global personality judgments. In this case, there would be no reason to assume that ratings given by domain experts or middle school students, who innately assess instructional features as they are related to their well-being, would have more predictive value compared to ratings of university students.

4.1 Reliability (research question 1)

As expected, we found thin slices ratings of instructional quality to be reliable within each of the three samples insofar as there was a high degree of agreement between the raters. Interestingly, experts agreed less on the level of single items compared to the other rater groups, while the ICC scores found on the level of scales are higher than for the other perspectives for classroom management and constructive support. Therefore, experts seem to offer slightly diverging opinions regarding single instructional practices (e.g. "The teacher helps his or her students when they have comprehension problems”). Yet, they are more consistent in their answers to items representing a single dimension. This could be due to their knowledge and familiarity with the distinctions of the Three Basic Dimensions, a framework of instructional quality that is well known among German-speaking educational researchers (Praetorius et al. 2018).

Findings from the Measures of Effective Teaching (MET) project indicate that including additional observers can increase the reliability of ratings of instructional features (Kane and Cantrell 2013). As such, high reliability of the thin slice ratings in this study may be in part due to the relatively large rater samples. The reliability of untrained observers’ thin slice ratings may depend on larger rater samples than ratings of instructional quality given by trained raters based on longer observations. Thus, there might be a trade-off of effort in training a few raters versus recruiting larger groups of raters. Determining the number of thin slices raters necessary to obtain reliable scores of instructional quality could be the subject of future studies.

4.2 Predictive validity (research question 2)

Our results suggest that ratings of instructional quality that are only based on the first impressions of untrained raters can predict how much students learn from different teachers. Thereby, we were able to replicate the findings from Begrich et al. (2017), which, to our knowledge, was the first study to investigate the practical usefulness of thin slices ratings in the context of instructional quality measures to explain differences in learning. We chose a quite conservative approach as we controlled for prior domain-specific knowledge, general scientific literacy as well as intelligence. Thus, thin slices ratings of instructional quality seem to also be predictive after controlling for these important student characteristics. Furthermore, our results show that the findings from Begrich et al. (2017) are generalizable to rater samples of experts in educational research as well as university students. Taken together, these findings suggest that thin slices ratings could serve as an approach for assessing the features of instructional quality that are relevant to learning. The final section discusses the practical implications of this finding.

4.3 Incremental validity (research question 3)

We tested three additional models to check for unique contributions of thin slices ratings by middle school students and experts to explain student learning. In each model, each group’s ratings of one dimension were introduced on the classroom level. Contrary to our expectation, domain experts’ ratings of cognitive activation did not explain a unique component of variance compared to ratings by university and middle school students. Instead, we found expert ratings of classroom management to explain a unique component of variance. Our previous findings might explain the lack of incremental validity of expert ratings of cognitive activation, as they reveal no overlap between thin slices ratings and trained raters’ assessments based on longer observations in regards to this dimension of instructional quality. In fact, this finding could be due to a lack of reliability in both thin slices ratings as well as the more systematic observer ratings—as a study by Praetorius et al. (2014) into the stability and variability of the three dimensions of instructional quality suggests. They find that classroom management and constructive support remain rather stable across lessons and, thus, can sufficiently be assessed based on observations of a single lesson. Reliable ratings of cognitive activation require at least nine lessons, however. The authors attribute this result to the dependence of cognitive activation on the content and type of the specific lesson that is observed. In addition, the three dimensions vary according to the degree to which they are manifested in observable behavior. While it is easy to detect teacher behavior such as establishing and communicating clear rules or giving supportive feedback to a child, the cognitive challenge inherent in a particular question or task is not immediately visible, for example. As thin slices ratings are based on cognitive processes that decode certain cues from the behavioral stream (Ambady et al. 2000), cognitive activation may not be a suitable construct for this measurement approach.

As experts' thin slices ratings of classroom management explain unique components of variance in students' test scores, it seems that they see more details relevant to this dimension of quality than university and middle school students when forming their impressions. Therefore, we found domain expertise to positively affect intuitive judgments of instructional quality. It should be noted, however, that the specific dimension that expert ratings were incrementally valid on was different than what we expected.

As anticipated, we found incremental validity of middle school students' thin slice ratings of constructive support. Constructive support is a quality dimension of instruction that is particularly important to emotional and motivational student outcomes. Therefore, it seems that middle school students are likely more sensitive to cues that are indicative for this quality dimension because of a special attunement. We consider this as additional support for the notion that it is possible to assess something specific to and diagnostic for instructional quality using thin slices ratings i.e. first impressions of untrained raters.

4.4 Limitations and future directions

This study contributes to the current state of research on assessing instructional quality. We interpret the finding that the incremental validity of thin slices ratings of instructional quality dimensions seems to be affected by domain expertise and attunement that are known to allow more accurate intuitive judgments in other domains as an indication that first impressions of untrained raters can be used to assess something specific to and diagnostic for instructional quality. Thus, thin slices ratings reflect more than global personality judgments. However, as our findings only indirectly address the question as to which information is decoded from 30-s observations, more research is necessary to clarify the degree to which thin slices ratings truly assess instructional quality. Future research should address this issue by checking for sensitivity of thin slices ratings to changes in instructional quality. A study could, for example, examine the improved outcomes of instructional quality as documented before and after an intervention by different sets of external observers based on classroom videos recorded at both points in time. Individual characteristics remain constant in our study, thus, differences between thin slices ratings of instructional quality could only be explained by the assessment of behavioral cues specific for these instructional features. Moreover, to compete with systematic ratings as a new measurement approach, thin slices ratings of instructional quality should be sensitive to improvements in the quality of instruction.

Another important question not addressed in the present study relates to the critical cues that lead the observers to their highly concordant judgments. It is reasonable to assume that these cues can be found in nuances of the teachers' (and students') non-verbal behavior (Ambady et al. 2000; Babad 2007). Differences in aspects of non-verbal behavior are thought to chronically communicate information about states, personality and the quality of interactions and relationships between two or more target persons. Thereby, they contribute to the first impressions of observers and form the basis of social judgments. One approach to examine the molecular behavioral cues responsible for certain thin slices judgments is called microanalysis (Babad 2007). Ambady and Rosenthal (1993), for example, applied microanalysis to check for a certain non-verbal profile of teachers that were rated favorable based on thin slices of their lecturing behavior. Even though their results did not reveal a specific pattern of non-verbal behaviors characterizing effective teachers we think it would be promising to apply microanalysis in our future studies to gain insight in the processes underlying thin slices ratings of instructional quality and potentially to distinguish them from processes (i.e. behavioral cues) underlying thin slices ratings of global personality traits. This could strengthen our argument that it is possible to assess something specific to instructional quality via thin slices ratings. Furthermore, we plan to use staged classroom videos in future studies to examine if thin slices ratings of instructional quality depend more on person related cues or if they are different in conditions where the same teacher (i.e. actor) shows high respectively low levels of practices indicating successful classroom management. The latter would be a hint of a decoding of instruction specific cues underlying thin slices ratings of classroom management.

Several other factors limit our results from being completely generalizable. For example, the video clips used in our study showed classroom instruction in elementary schools. Therefore, it is questionable whether our results also apply to instruction in other school forms. Furthermore, it would be interesting to examine if the effects of students' attunement to cues of constructive support are even clearer in thin slices ratings of students observing instruction in their own grade. Second, the classroom videos all show the implementation of the same standardized instructional unit. Since teachers are observed in similar settings and situations, differences in the quality of their instruction may be especially visible. Further investigations are needed to determine whether instructional quality is also assessable based on short observations that show teachers within differing instructional contexts and settings. Third, we have a heterogeneous sample of experts as well as no information about how much constructive support is emotionally relevant to the specific sample of middle school students. It may be promising to directly assess these characteristics in both rater groups to examine the effects of expertise and attunement on the accuracy of thin slice ratings. Finally, it could be worthwhile to investigate how other rater characteristics besides domain expertise and attunement affect the predictive value of thin slices ratings of instructional quality. For example, there is evidence suggesting that social perceivers with higher interpersonal orientation and social adjustment can more accurately judge others than people who are less interpersonal oriented and socially adjusted (see e.g. Davis and Kraus 1997). It is important to examine the effects of such characteristics on the reliability and validity of thin slices ratings to find the most suitable population of raters for assessing quality of instruction via first impressions.

4.5 Practical implications

Overall, we found that short glimpses of teacher behavior can serve as the basis for overall judgments about that teacher's general instructional quality. Moreover, we found that first impressions about quality of instruction are predictive for students' learning across different groups of raters. Our study of thin slices ratings as a potentially economical approach for assessing instructional quality stands within the context of a growing stream of research devoted to the measurement of instructional quality and teacher effectiveness using ratings by external observers (see e.g. Good and Lavigne 2015; Kane et al. 2012, 2014; Pianta and Hamre 2009; Praetorius et al. 2012; Strong et al. 2011). In light of the complex and time-consuming rating schemes typically used to assess instructional quality in video studies (e.g. Kane and Cantrell 2013; Pianta and Hamre 2009), these results are surprising and perhaps also provocative (see Good and Lavigne 2015 and Gargani and Strong 2015 for an intensive debate on that matter).

It is important to note that more evidence regarding the validity of these ratings is needed before we can make actual suggestions about their practical usefulness. It is also important to consider different occasions and purposes that require the measurement of instructional quality to exclude potential areas of application. One such instance is assessing the quality of instruction for the purpose of professional development, i.e. using diagnostic information to improve teachers' instruction. The Measures of Effective Teaching project (MET; see Kane et al. 2014) has invested enormous effort into evaluating different observational instruments that suit that purpose. Results gathered by this research project suggest that high-quality classroom observations require multiple observations as well as scores by several highly trained and certified observers to be averaged. Furthermore, instruments must assess a variety of instructional practices to give teachers productive feedback by taking into account certain strengths and weaknesses (Good and Lavigne 2015). As we use only six items to assess features of instruction that are relevant to learning, this approach is clearly unsuitable for providing feedback and detailed diagnostics on the quality of individual teachers' instruction. This is especially true when assessing instructional quality serves as a basis for high-stake decisions that could affect a teacher's career, for instance.

If future studies find further evidence for the validity of thin slices ratings, however, a potential area of application could be in assessing instructional quality across larger samples of teachers. Within the context of research on instruction thin slices ratings could help to explain differences in student outcomes. Especially in large-scale assessments where common observation measures are often too expensive and time-consuming to be applied to the full sets of classes, thin slices ratings of instructional quality could be a complementary research tool. Furthermore, thin slices ratings may serve as a screening instrument that helps to find cases of successful or struggling teachers, who require a more detailed diagnostic when it comes to professional development. Therefore, we find it promising to continue exploring the potential of thin slices ratings for assessing instructional quality.

Acknowledgements Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

Ethical approval The study was carried out in accordance with the regulations of the department's ethic committee. The committee waives a detailed examination of an empirical study if the following conditions are met: It can be firmly assumed that participation in the study will not cause any conceivable physical or mental harm or discomfort for the participants that exceed their every-day experiences. 2. The study is based on data anonymized at the source. Rating questionnaires were anonymous. 3. The study is partly based on archival material (classroom videos, student data), for which confidentiality is ensured, that the attribution of data to a specific person is not possible. All these conditions are met in our study, thus no formal ethical approval statement was issued.

References

Ambady, N. (2010). The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry, 21*(4), 271–278.

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Towards a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology, 32,* 201–271.

Ambady, N., & Grey, H. F. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology, 83,* 947–961.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teachers evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431–441.

Babad, E. (2005). Guessing teachers differential treatment of high and low achievers from thin slices of their public lecturing behavior. *Journal of Nonverbal Behavior, 29*(2), 125–134.

Babad, E. (2007). Teachers' nonverbal behavior and its effects on students. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 201–261). New York: Springer.

Babad, E., Bernieri, F., & Rosenthal, R. (1991). Students as judges of teachers' verbal and nonverbal behavior. *American Educational Research Journal, 28*(1), 211–234. https://doi.org/10.3102/00028312028001211.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., & Jordan, A. (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.

Begrich, L., Fauth, B., Kunter, M., & Klieme, E. (2017). Wie informativist der erste Eindruck? Das Thin-Slices-Verfahren zur videobasierten Erfassung des Unterrichts. [How informative is a first impression? The thin slices procedure for the video-based assessment of instruction.]. *Zeitschrift für Erziehungswissenschaft, 20*(1), 23–47. https://doi.org/10.1007/s11618-017-0730-x.

Begrich, L., Kuger, S., Klieme, E., & Kunter, M. (2019). *At first glance—Using the thin slices technique to assess instructional quality* (manuscript submitted for publication).

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology, 86*(4), 599–614.

Bos, W., Bonsen, M., Baumert, J., Prenzel, M., Selter, C., & Walther, G. (Eds.). (2007). *TIMSS 2007—Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. [TIMSS 2007—Mathematical and scientific competencies of primary school children in Germany in an international comparison].* Münster: Waxmann.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (pp. 1–36). Hillsdale: Erlbaum.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminate validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*(5), 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004.

Carrère, S., & Gottman, J. M. (1999). Predicting divorce among newlyweds from the first three minutes of a marital conflict discussion. *Family Process, 38*(3), 293–301.

Cianciolo, A. T., Matthew, C., Sternberg, R. J., & Wagner, R. K. (2006). Tacit knowledge, practical intelligence, and expertise. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 613–632). New York: Cambridge University Press.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? [Quality of instruction: A matter of perspective?].* Münster: Waxmann.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023. https://doi.org/10.1037/0022-3514.92.6.1006.

Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The interpersonal perception task. *Journal of Nonverbal Behavior, 13,* 223–245.

Dane, E., & Pratt, M. G. (2007). Exploring intuition and its role in managerial decision making. *Academy of Management Review, 32,* 33–54.

Dane, E., Rockmann, K. W., & Pratt, M. G. (2012). When should I trust my gut? Linking domain expertise to intuitive decision-making effectiveness. *Organizational Behavior and Human Decision Processes, 119*(2), 187–194. https://doi.org/10.1016/j.obhdp.2012.07.009.

Davis, M. H., & Kraus, L. A. (1997). Personality and empathic accuracy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 144–168). New York: Guilford.

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., et al. (2015). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research, 108*(5), 358–370.

Dreyfus, H. L., & Dreyfus, S. E. (2005). Expertise in real world contexts. *Organization Studies, 26,* 779–792.

Emmer, E. T., & Evertson, C. M. (2013). *Classroom management for middle and high school teachers*. Boston: Pearson.

Emmer, E. T., & Stough, L. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36,* 103–112.

Ericsson, K. A., & Charness, N. (1994). Expert performance. Its structures and acquisition. *American Psychologist, 49,* 725–747.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014a). Grundschulunterricht aus Schüler-, Lehrer-und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg [Instruction in elementary school from the perspective of students, teachers, and observers: Correlations and prediction od learning success]. *Zeitschrift für Pädagogische Psychologie*, *28*(3), 127–137.

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

Feeley, T. H. (2002). Comment on Halo effects in rating and evaluation research. *Human Communication Research, 28*(4), 578–586. https://doi.org/10.1111/j.1468-2958.2002.tb00825.x.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Applied Psychology, 66*(2), 127–148.

Fiske, S., & Neuberg, S. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in experimental social psychology* (23rd ed., pp. 1–75). San Diego: Academic Press Inc.

Fowler, K. A., Lilienfeld, S. O., & Patrick, C. J. (2009). Detecting psychopathy from thin slices of behavior. *Psychological Assessment, 21*(1), 68–78.

Friedman, J. N. W., Oltmanns, T. F., & Turkheimer, E. (2007). Interpersonal perception and personality disorders: Utilization of a thin slice approach. *Journal of Research in Personality, 41*(3), 667–688.

Gargani, J., & Strong, M. (2014). Can we identify a successful teacher better, faster, and cheaper? Evidence for innovating teacher observation systems. *Journal of Teacher Education, 65*(5), 389–401. https://doi.org/10.1177/0022487114542519.

Gargani, J., & Strong, M. (2015). Response to "Rating teachers cheaper, faster, and better: Not so fast": It's about evidence. *Journal of Teacher Education, 66*(4), 395–401. https://doi.org/10.1177/0022487115587110.

Gibson, J. J. (1979). *The ecological to visual perception*. Boston: Houghton-Mifflin.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine, 86*(10), 1–7.

Good, T. L., & Lavigne, A. L. (2015). Rating teachers cheaper, faster, and better: Not so fast. *Journal of Teacher Education, 66*(3), 288–293. https://doi.org/10.1177/0022487115574292.

Hardy, I., Hertel, S., Kunter, M., Klieme, E., Warwas, J., & Büttner, G. (2011). Adaptive Lerngelegenheiten in der Grundschule: Merkmale, methodisch-didaktische Schwerpunktsetzungen und erforderliche Lehrerkompetenzen. [Adaptive learning opportunities in elementary school: Characteristics, methodological-didactic focus, and required teacher competencies]. *Zeitschrift für Pädagogik, 57*(6), 819–833.

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology, 98,* 307–326.

Hardy, I., Kleickmann, T., Koerber, S., Mayer, D., Möller, K., Pollmeier, J., et al. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter. [The modeling of scientific competencies in primary school age]. In E. Klieme, D. Leutner, & M. Kenk (Eds.), *Kompetenzmodellierung* (pp. 115–125). Weinheim: Beltz.

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Holleran, S. E., Mehl, M. R., & Levitt, S. (2009). Eavesdropping on social life: The accuracy of stranger ratings of daily behavior from thin slices of natural conversations. *Journal of Research in Personality, 43*(4), 660–672.

Holzberger, D., Praetorius, A. K., Seidel, T., & Kunter, M. (2019). Identifying effective teachers: The relation between teaching profiles and students' development in achievement and enjoyment. *European Journal of Psychology of Education, 34*(4), 801–823. https://doi.org/10.1007/s10212-018-00410-8.

Jung, M. F. (2016). Coupling interactions and performance: Predicting team performance from thin slices of conflict. *ACM Transactions on Computer–Human Interactions, 23*(3), 1–32. https://doi.org/10.1145/2753767.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, and Giroux.

Kane, T. J., & Cantrell, S. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study.* Retrieved January 8, 2019 from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Kane, T., Kerr, K., & Pianta, R. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the measures of effecting project*. San Francisco: Wiley.

Kane, T., Staiger, D., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., et al. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Technical report. Seattle: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project.

Klieme, E. (2006). Empirische Unterrichtsforschung: Aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. [Empirical research on teaching: Current trends, theoretical background and subject-specific findings]. *Zeitschrift für Pädagogik, 52,* 765–773.

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

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*(2), 284–308. https://doi.org/10.1037/0033-295x.103.2.284.

Künsting, J., Neuber, V., & Lipowsky, F. (2016). Teacher self-efficacy as a long-term predictor of instructional quality in the classroom. *European Journal of Psychology of Education, 31*(3), 299–322. https://doi.org/10.1007/s10212-015-0272-7.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. https://doi.org/10.1037/a0032583.

Kunter, M., & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. [The model of instructional quality in COACTIV: A multi-criterial analysis]. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Professionelle Kompetenz von Lehrkräften—Ergebnisse des Forschungs programms COACTIV* (pp. 85–113). Münster: Waxmann.

Lambert, N., Seth, M., & Frank, F. (2014). Thin slices of infidelity: Determining whether observers can pick out cheaters from a video clip interaction and what tips them off. *Personal Relationships, 21*(4), 612–619. https://doi.org/10.1111/pere.12052.

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

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings in multilevel modelling. *Contemporary Educational Psychology*, *34*, 120–131.

Mainhard, T., Wubbels, T., & Brekelmans, M. (2014). The role of the degree of acquaintance with teachers on students' interpersonal perceptions of their teacher. *Social Psychology of Education, 17*(1), 127–140. https://doi.org/10.1007/s11218-013-9234-6.

Manson, J. H., Gervais, M. M., & Bryant, G. A. (2018). General trust impedes perception of self-reported primary psychopathy in thin slices of social interaction. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0196729.

Marzano, R., Marzano, J., & Pickering, D. (2003). *Classroom management that works*. Alexandria: Association for Supervision and Curriculum Development.

McDonald, R. P. (1970). The theoretical foundations of principle factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23,* 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahawah, NJ: Erlbaum.

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology, 71,* 39–44.

Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus user's guide* (7th ed.). Los Angeles: Muthén & Muthén.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48,* 630–635.

Oltmanns, T. F., Friedman, J. N. W., Riedler, E. R., & Turkheimer, E. (2004). Perceptions of people with personality disorders based on thin slices of behavior. *Journal of Research in Personality, 38*(3), 216–229.

Oltmanns, T. F., & Turkheimer, E. (2006). Perceptions of self and others regarding pathological personality traits. In R. F. Krueger & J. L. Tackett (Eds.), *Personality and psychopathology* (pp. 71–111). New York: Guilford.

Patterson, M. L., & Stockbridge, E. (1998). Effects of cognitive demand and judgment strategy on person perception accuracy. *Journal of Nonverbal Behavior, 22,* 253–263.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. https://doi.org/10.3102/0013189x09332374.

Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., et al. (2015). Judging students' self-concepts within 30 s? Investigating judgement accuracy in a zero-acquaintance situation. *Learning and Individual Differences, 37,* 231–236. https://doi.org/10.1016/j.lindif.2014.11.015.

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

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

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

Pretsch, J., Flunger, B., Heckman, N., & Schmitt, M. (2013). Done in 60 s? Inferring teachers subjective well-being from thin slices of non-verbal behavior. *Social Psychology of Education, 16*(3), 421–434.

Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C., & Reusser, K. (2007). Structure as a quality feature in mathematics instruction of the learning environment vs. a structured presentation of learning content. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG priority programme* (pp. 101–120). Münster: Waxmann.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2000). *Teachers, schools, and academic achievement*. Working Paper W6691. Cambridge: National Bureau of Economic Research.

Rosenthal, R. (1987). *Judgement studies*. New York: Cambridge 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.

Simon, H. A. (1987). Making management decisions: The role of intuition and emotion. *Academy of Management Executive, 1*(1), 57–64.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Science, 23,* 645–726.

Strong, M., Gargani, J., & Hacifazlioglu, Ö. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education, 62*(4), 367–382.

Tskhay, K. O., Zhu, R., & Rule, N. O. (2017). Perceptions of charisma from thin slices of behavior predict leadership prototypicality judgments. *The Leadership Quarterly, 28*(4), 555–562. https://doi.org/10.1016/j.leaqua.2017.03.003.

Visser, D., & Matthews, J. D. L. (2005). The power of non-verbal communication: Predicting job performance by means of thin slices of non-verbal behaviour. *South African Journal of Psychology, 35*(2), 362–383. https://doi.org/10.1177/008124630503500212.

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. [Swizz math instruction from a students' point of view and from the perspective of high-inferent observer ratings]. In K. Reusser, C. Pauli, & M. Waldis (Eds.), *Unterrichtsgestaltung und Unterrichtsqualität* (pp. 171–208). Münster: Waxmann.

Walton, K. M., & Ingersoll, B. R. (2016). The utility of thin slice ratings for predicting language growth in children with autism spectrum disorder. *Autism, 20*(3), 374–380. https://doi.org/10.1177/1362361315584465.

Weiß, R. H. (2006). *CFT 20-R. Grund intelligenz test Skala 2—Revision [CFT 20-R. Basic intelligence test scale 2—Revision]*. Göttingen: Hogrefe.

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education*, *19*(3), 409–427.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Lukas Begrich is a scientific co-worker at the Department of Educational Psychology of the Goethe University Frankfurt in Germany. He recently finished his dissertation in which he examined the possibility of assessing instructional quality based on thin slices ratings.

Benjamin Fauth is Head of the Department for Empirical Educational Research at the Institute for Educational Analysis (IBBW) in Stuttgart, Germany and Extraordinary Professor at the University of Tübingen, Germany. His research focuses on teaching quality, professional competence of teachers and questions of applied evaluation research.

Mareike Kunter is a Professor of Educational Psychology at Goethe University, Frankfurt. Her research focuses on professional competence of teachers, teacher career development, classroom instruction, and motivation at school.