Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch Vol. 12 No. 3 (2024) 69 - 98
ISSN 2295-3159

The Underlying Cognitive Processes of Thin Slices Judgments on Teaching Quality

Konstantin Vinokic [1, 2], Lukas Begrich [2], Mareike Kunter [2] & Susanne Kuger [3]

[1] Institute for Educational Analysis Baden-Württemberg (IBBW), Germany
[2] DIPF, Germany
[3] German Youth Institute (DJI),Germany

Article received 29 December 2023 / Article revised 19 September 2024 / Accepted 2 October 2024 / Available online 25th October 2024

Abstract

Thin slices ratings (i.e., ratings based on first impressions) have yielded intriguingly accurate results in various domains. Among other, researcher have applied the thin slices technique to assess instructional quality, showing that teacher-student interactions can be reliably inferred by just very short snippets of classroom instruction. The accuracy of thin slices ratings is often explained by dual process theories of social cognition, whereby System 1 refers to an intuitive and fast way of processing, while System 2 denotes a more reflective and analytical way of processing. System 1 is considered the cognitive foundation of thin slices ratings. The central aim of the present study was to understand the underlying cognitive processes shaping the impression formation of thin slices raters of teaching quality. Therefore, an unconventional and innovative research design was required to gain insights into the cognitive "black box" of thin slices raters by examining their verbal data. In an exploratory mixed method research design, we set up Cognitive Laboratories with two different rating situations. In a thin slices rating situation, participants rated instructional quality based on short classroom videos (30 seconds). Participants in a long-video rating situation rated instructional quality based on longer classroom videos (10 minutes). We collected, coded and statistically analyzed participants' verbal reports regarding their rating processes. The findings suggest that thin slices ratings evolve primarily based on typical processes of System 1 and not on those of System 2. For instance, thin slices ratings are associative and tend to be rather negative than positive. Moreover, an initially formed impression tends to remain stable and is resistant to alteration. Ratings of instructional quality based on longer videos rely on both cognitive systems, with System 2 possibly modifying an initial judgment. Thus, our study does not only explain the cognitive processes underlying the thin slices ratings, but additionally provides valuable insights into the processes occurring in conventional rating settings.

Keywords: thin slices technique; instructional research; dual process theories of social cognition; first impressions; rater cognition

Konstantin Vinokic, Rostocker Str. 6, 60323 Frankfurt am Main, Germany, k.vinokic@dipf.de (alternative Email address: vinokic@gmx.de) *DOI:* https://doi.org/10.14786/flr.v12i3.51421

Video-based analysis of teaching quality is an important source of data collection in educational research (de Freitas, 2015; Hannafin et al., 2010; Petko et al., 2003). The production of classroom videos, the rater training programs, and the observational rating procedure of video lessons are effortful, time-consuming and costly (Gargani & Strong, 2014; Kane & Staiger, 2012; Murphy & Hall, 2021). Thus, research might benefit from a more economical rating approach that is less resource-demanding and still produces reliable and valid data. The so-called thin slices technique (i.e., ratings based on first impressions) could be such a less effortful and more economical solution and might extend the repertoire of data collection methods for instructional research (Ambady et al., 2000; Murphy & Hall, 2021).

In a seminal work on first impressions, Asch (1946) systematically examined the cognitive processes that underlie impression formation. Although individuals can form accurate impressions of faces within 100 milliseconds (Willis & Todorov, 2006), first impressions are typically considered to develop over the course of five minutes (e.g., Wood, 2014). The thin slices technique investigates the accuracy of judgments based on first impressions. A thin slice is defined as an excerpt of dynamic behavior less than five minutes long (Ambady et al., 2000). Studies, applying the thin slices technique in the context of teaching, have shown that it can yield accurate results in terms of reliability and validity (Ambady & Rosenthal, 1993; Babad, 2005; Begrich et al., 2021; Sokolovic et al., 2021; Vinokic et al., 2024). In contrast to the ample evidence that thin slices ratings work, there is far less research on why they work so well. A potential explanation for the surprising accuracy of thin slices judgments might be a universal human cognitive principle that is described by dual process theories of social cognition. Dual process theories distinguish between two cognitive systems of information processing. System 1 operates fast, autonomously and intuitively, whereas System 2 functions slowly, reflectively, analytically and consciously (Kahneman, 2003, 2011; Gawronski et al., 2024). Theoretically, it has been argued that thin slices ratings operate with cognitive processes associated with System 1 (Wood, 2014), however this claim has only rarely been examined empirically (e.g., Ambady, 2010). Consequently, the present study examines whether participants' cognitive processes during their thin slices ratings of instructional quality are indeed in accordance with System 1 of dual process theories. In order to gain insights into the black box of the underlying cognitive processes of thin slices ratings of teaching quality, the study pursued a rather unconventional and innovative approach. We developed Cognitive Laboratories with an exploratory, mixed method research design, using think aloud protocols and guided interviews. To date, nothing is known about the cognitive processes shaping the impression formation of thin slices raters assessing teaching quality. The present study aims to explore various cognitive mechanisms, including cognitive biases, strategies as well as the roles of intuition and associations in the context of rapid information processing when assessing teaching quality. Additionally, the results may have implications for understanding the judgment formation of conventional raters of teaching quality. Since typical cognitive processes associated with first impressions also appear to affect the judgment formation of conventional raters, the results of the present study provide valuable insights to improve the partially insufficient reliability of conventional raters (Praetorius et al., 2012; White & Ronfeldt, 2024).

1. Theoretical Background

1.1 The Thin Slices Technique

Day-to-day impressions and judgments about others are often formed rapidly, consciously and intuitively based on very few information. The so-called thin slices research technique was developed

to test the accuracy of ratings based on scarce information. Thin slices judgments rely on first impressions of observers who are not interacting with target persons (Ambady et al., 2000; Wood, 2014). A thin slice as a brief excerpt of dynamic information sampled from the behavioral stream less than five minutes long. The source of information might be audio only, video only or audiovisual (Ambady et al., 2000). Research has demonstrated thin slices judgments of naive or untrained raters to be highly accurate in terms of reliability (raters agree on their judgments) and validity (ratings of judges correlate with external criteria) (Begrich et al., 2020; Fowler et al., 2009; Tackett et al., 2015; Pretsch et al., 2013).

The thin slices technique had been applied in various research fields such as social psychology (Jung, 2016), personality psychology (Holleran et al., 2009) or in the clinical context (Rimondini et al., 2019). For example, Lambert et al. (2014) found that raters who had watched 3-4-minute videos of unknown couples were able to correctly identify those persons who had cheated on their partner. Visser and Matthews (2005) showed that ratings of observers who had watched 30-second clips of call center operators' nonverbal behavior could successfully predict the operators' job performances as rated by managers and customers. Borkenau et al. (2004) applied the thin slices technique to examine the accuracy of Big Five personality traits and intelligence. The ratings were based on 3-minute videos that showed target persons engaging in various activities. They found significant correlation between thin slices judgments and the outcomes on standardized intelligence tests as well as between thin slices judgments and personality reports of familiar persons.

The thin slices technique has also been applied to educational research. Ambady and Rosenthal (1993) found significant correlations between students' end-of-semester evaluation of teachers and thin slices ratings (30 seconds) of teachers' personality by naive raters. Even shorter slices (6 seconds, 15 seconds) were strongly related to the criterion variables. Babad (2005) demonstrated that high school students could significantly predict differential behavior of unfamiliar teachers toward high- and low-achieving students based on 10-second silent clips. In the context of early childhood education and care (ECEC), trained thin slices raters could reliably and validly assess the quality of interaction between teachers and children even though the teachers were wearing facial masks (Sokolovic et al., 2021; Vinokic et al., 2024). Begrich et al. (2017, 2020, 2021) applied the thin slices technique in order to assess instructional quality based on 30-seconds classroom video clips. They examined the accuracy of thin slices ratings given by naive, untrained raters, demonstrating that thin slices ratings correlated highly among raters (reliability), correlated substantially with the ratings of trained observers based on full lessons (convergent validity), captured distinctively three different dimensions of instructional quality (construct validity) and predicted students' outcomes (predictive validity). Overall, the body of thin slices research indicates that trained and untrained raters highly agree in their judgments, and these judgments correlate with external criteria. This raises the question of how such highly accurate judgments are formed on the basis of scarce information. What are the underlying cognitive processes of thin slices ratings responsible for this remarkable accuracy?

1.2 Dual Process Theory to Explain the Accuracy of Thin Slices Judgments

In cognitive psychology, dual process theories try to explain social cognition with two systems: System 1 and System 2 (Evans, 2006, 2019; Kahneman, 2011; Milli et al., 2021; Stanovich & West, 2000). System 1 is supposed to operate automatically, quickly, emotionally and associatively, with no or little effort, and without a sense of voluntary control (Kahneman, 2011). It operates unconsciously, rapidly and intuitively with high capacity (Stanovich, 2009), relying on processes that are considered to be evolutionarily old and more directly tuned to ancient reproductive goals (Evans, 2008; Pennycook, 2017; Stanovich, 2009). System 1 processes operate autonomously and holistically, do not require working memory and are thought to be domain-specific (Bellini-Leite, 2018; Evans & Stanovich, 2013; Stanovich & West, 2000; Nisbett et al., 2001). From an evolutionary perspective, first impressions are easily pulled to the negative end of an evaluative dimension. A negative impression may be formed on the basis of very few information and a positive impression may require a greater amount of information (Ambady & Skowronski, 2008; Nesse, 2005). According to the smoke detector principle, the fitness

costs for a false alarm (i.e., an incorrect negative first impression) are lower than a missed alarm in case of danger (i.e., an incorrect positive first impression; Nesse, 2005).

In contrast, System 2 describes a conscious, slow, deliberate, analytical and reflective way of processing (Evans, 2006; 2008). System 2 processes are considered to be evolutionary young, domain-general, capacity-limited and rule-based (Pennycook, 2017). System 2 requires attention, is often associated with concentration and reasoning, and allocates attention to demanding, effortful activities, such as complex computation (Kahneman, 2011). System 2 operates in a rule-based manner, compares objects across several attributes, makes deliberate choices between options (Kahneman, 2011), and encompasses the processes of analytic intelligence (Stanovich & West, 2000).

The terminology System 1 and System 2 has become popular; however, different terminologies are debated. Stanovich et al. (2014) and Evans (2018; 2019) propose to refer to these two systems as Type 1 and Type 2. Bellini-Leite (2018) discusses various terminologies, like systems, types, clusters or modes. Besides the most prominent dual process theory, namely the *default-interventionists dual process theory*, other accounts exist, such as the *parallel dual process theory* or the *unisystem model*. The *default-interventionists dual process theory,* advocated by e.g., Kahnemann (2011) and Evans (2006, 2019), proposes that System 1 is the default and can be overridden by System 2. System 2 might take over and is able to overrule the freewheeling associations and impulses of System 1 (Kahneman, 2011). How these two systems precisely interact is still under debate. Although Mugg (2015) does not support the default-interventionist approach, he outlines the literature and claims that Type 1 and Type 2 processes are generated at different times: At first, Type 1 and then, under certain conditions, Type 2 processes are activated. Type 2 processing intervenes on Type 1 responses rather than generating an independent response on its own (Mugg, 2015). Overriding System 1 cognitions is more likely to occur in persons with higher analytic intelligence because they are more prone to produce responses that are epistemically and instrumentally rational (Stanovich, 2009). The *parallel dual process theory* claims two reasoning systems operating in parallel and competing against each other, with System 1 being faster than System 2 (Mugg, 2015; Sloman, 1996). Stanovich (2009) instead proposes a *tripartite mind* differentiating System 2 into the algorithmic and reflective mind and referring to System 1 as the autonomous mind. In contrast to dual process theories, *one-system accounts* or *unisystem models* postulate only one reasoning system, which operates along a continuum (De Neys, 2021; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Mugg, 2015).

Although impressions based on System 1 processes can be biased (for an overview see Wood, 2014), the surprisingly accurate thin slices ratings are thought to rely predominantly on System 1 processes due to the speed of judgment formation and the limited involvement of cognitive resources (Ambady et al., 2000; Murphy & Hall, 2021; Wood, 2014). However, to our knowledge, there are only a few studies (Ambady, 2010) that have explicitly and globally examined this assumption. This raises the question of whether the underlying cognitive processes of thin slices ratings of teaching quality feature characteristics that are typical for System 1?

1.3 Instructional Quality

There is wide consensus that instructional quality is a key determinant of students' achievement (Decristan et al., 2016; Hattie, 2023). Klieme described a three-dimensional structure of instructional quality and its relevance for explaining students' outcomes (Fauth et al., 2024; Klieme et al., 2001, 2009; Kunter et al., 2005; Trautwein et al., 2022). These so-called three basic dimensions of instructional quality are well established, especially in German-speaking countries (Kuger et al., 2017; Praetorius et al., 2018). The first dimension, classroom management, refers to the teacher's ability to sustain an orderly and functioning classroom setting, which involves aspects such as classroom discipline, effective handling of disruptions and clarity of rules. Further key features of classroom management are smooth transitions between tasks and effective time-on-task learning (Decristan et al., 2016; Doyle, 2006; Kuger et al., 2017; Marzano & Marzano, 2003). The second basic dimension, cognitive activation, emphasizes the teacher's ability to engage students in higher-order thinking by encouraging them to

actively process and reflect on the learning material, rather than passively receive information. Cognitively challenging questions can create cognitive conflicts, which can lead to a deeper understanding of concepts (Baumert et al., 2010; Decristan et al., 2016; Lipowsky et al., 2009). The third basic dimension, constructive support, refers to a teacher's supportive behavior, such as caring about individual needs, stimulating personalized learning, motivating students, and providing constructive feedback (Decristan et al., 2022; Kuger et al., 2017; Kunter & Voss, 2011; Praetorius et al., 2018).

Instructional quality is usually assessed either by surveys taken from teachers and students or by trained external observers, applying a complex coding system (Janik & Seidel, 2013; Kunter & Baumert, 2006; Pianta & Hamre, 2009). However, it is doubtful whether student ratings provide accurate results due to students' lack of didactic knowledge, their involvement in the instructional process, and confounding factors such as teacher popularity (Aleamoni, 1999; Clausen, 2002). However, recent empirical evidence indicates that students can validly assess instructional quality, albeit depending on the dimension (Göllner et al., 2021). Teachers' self-reports have shown low agreement with other approaches of assessment as well as low predictive validity regarding student outcomes (Clausen, 2002; Desimone et al., 2010; Wagner et al., 2016). In the light of these constraints, ratings from external observers based on classroom videos are often seen as the most valid way to receive information about instructional processes (Helmke, 2014; Pianta & Hamre, 2009; Praetorius et al., 2012). However, ratings from external observers are highly resource-demanding in terms of money and time (Helmke, 2014; Janik & Seidel, 2013).

Against this backdrop of the high costs of conventional video rating procedures, Begrich et al. (2017, 2020, 2021) examined whether the economical thin slices technique can yield accurate results in assessing instructional quality. Begrich and colleagues (2017, 2020, 2021) found evidence that thin slices raters highly agree in their judgments and that thin slices ratings can even predict students' short-term learning outcomes. Therefore, the thin slices technique might bare the potential to become an economic complement to the so far established approaches of data collection in instructional research.

2. Research Aims

The thin slices technique is based on first impressions of untrained raters. It was successfully applied in various domains and proved to be very accurate in terms of reliability and validity (e.g., Lambert et al., 2014; Visser & Matthews, 2005). Moreover, thin slices ratings appear to deliver sound results in the context of instructional research (Begrich et al., 2017, 2020, 2021). The intriguing precision of thin slices ratings seems to be counterintuitive, given the complex and contextualized nature of teaching (Berliner, 2005). Thus, not surprisingly, researchers have challenged the validity of thin slices ratings of instructional quality. However, arguing from the perspective of the dual process theory of social cognition (Evans, 2008; Stanovich et al., 2014), it is yet conceivable that first impressions of teaching quality may convey useful information. Thin slices ratings presumably rely on System 1 processes, i.e., intuitive, holistic and automatized cognitive processes that seem to be able to sort and evaluate complex information rapidly, as shown in many other areas of human interaction (Murphy & Hall, 2021; Wood, 2014). The activation of System 1 has been named as a reason for the surprisingly high accuracy of thin slices ratings (Ambady et al., 2000; Wood, 2014). However, this claim has only rarely been explicitly and globally examined empirically (Ambady, 2010).

The present study thus explores a potential link between cognitive processes underlying thin slices judgments of instructional quality and processes of System 1, as described in dual process theories of social cognition (e.g., Ambady & Skowronski, 2008; Kahneman, 2011; Stanovich 2009). In detail, the study focuses on the following two research questions (RQs).

Research Question 1: Do the reported mental processes underlying thin slices ratings of instructional quality resemble typical System 1 functioning? Of particular interest is whether typical

System 1 characteristics can be identified in the judgmental processes reported by thin slices raters. Based on the literature, we expect System 1 to serve as the underlying foundation of thin slices ratings.

Research Question 2: Are the cognitive processes underlying thin slices ratings not only similar to typical System 1 processes, but also substantially dissimilar to typical System 2 processes? Beyond finding proof that System 1 processes are the foundation of thin slices ratings, a stronger proof of concept would involve discriminant evidence showing the dissimilarity between cognitive processes during thin slices ratings and typical System 2 processes. We expect to find less evidence for typical processes of System 2 in the verbal reports of thin slices raters than in the verbal reports of raters whose judgments are based on a traditional systematic video rating approach (i.e., relying on more information).

3. Method

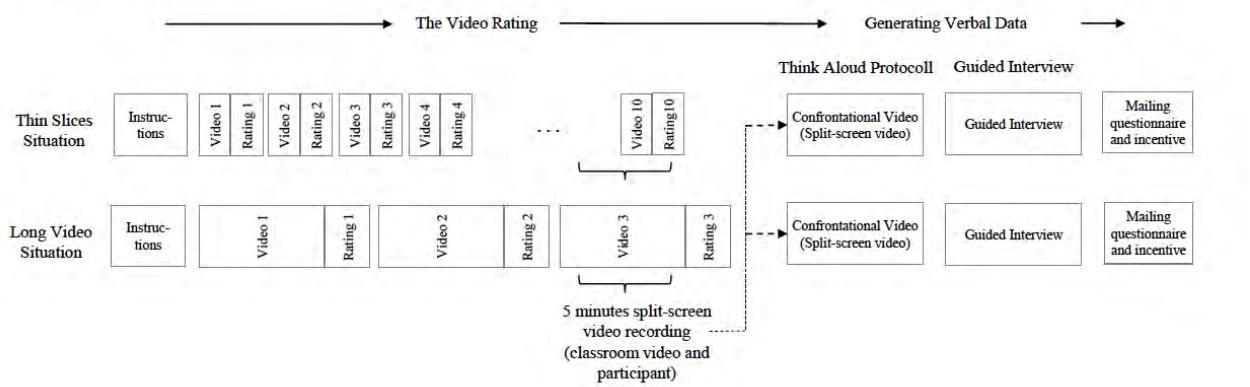
3.1 Overview

In the present study, a mixed method research design was pursued (McKim, 2017; Schoonenboom & Johnson, 2017; Tashakkori & Creswell, 2007) by conducting Cognitive Laboratories (Hyytinen et al., 2014; Leighton, 2017) with two different rating situations. In one rating situation, participants assessed instructional quality based on 30-second classroom videos, which is considered a typical thin slices rating situation (SliceS). In the other rating situation, participants assessed instructional quality based on 10-minute classroom videos (LongS; see section 3.2). We obtained qualitative data through think aloud protocols and guided interviews and analyzed them by applying the coding method for qualitative research (Saldaña, 2009). We combined an inductive and deductive coding approach. The coding aimed to identify patterns in the verbal data indicating underlying cognitive processes. The qualitative data were quantified by generating frequency counts of the codes, enabling us to detect differences between the two rating situations (Kawulich, 2004; Sandelowski et al., 2009).

3.2 Design and Procedure

Prior to the main study, we conducted trials with eight volunteers. During the trial, we tested and refined the technical procedure and the interview. Simultaneously, we initiated the analysis process by trying to identify patterns in the data. Due to the global COVID-19 pandemic, we conducted the trial and the main study completely online in one-to-one sessions. The internet connection was good or acceptable. The study consisted of three successive phases (Fig. 1). First, participants watched the videos and completed the rating questionnaire according to their rating situation (SliceS or LongS). During this phase, each participant was videotaped (see 3.2.1). Second, we confronted participants with their recording (split-screen video) and asked them to think aloud (see 3.2.2). Third, a guided interview was conducted (see 3.2.2). We audio-recorded the think aloud protocols and the guided interviews, giving us two relevant sources of data: the think aloud protocols and the guided interviews. The audio recordings were transcribed according to the system of transcription developed by Dresing and Pehl (2012).

Figure 1. Research Design



3.2.1 The Rating Situations

Two different rating situations were implemented (Fig. 1): a typical thin slices rating situation and a typical conventional rating situation. In the typical thin slices rating situation (SliceS), participants watched 10 classroom videos, each lasting 30 seconds, featuring 10 different teachers (Fig. 1). After each video, participants completed a rating questionnaire on classroom quality. The six-point Likert scale interactive questionnaire is based on Begrich et al. (2017) and consisted of six items representing the three basic dimensions of instructional quality (Praetorius et al., 2018). The first of these ten classroom videos served as a trial. Since we intended to set up a typical thin slices situation, participants were instructed to rely on their first impressions while rating the teachers' behavior, and they were given only 30 seconds to fill in the rating questionnaire (e.g., Begrich et al., 2017, 2020, 2021).

In the LongS, participants watched three classroom videos of 10 minutes in length each. After each video, they completed the identical rating questionnaire (Fig. 1). To approximate the rating procedure of a typical classroom observation study (Hardy et al., 2011; DESI-consortium, 2008), participants were instructed to think about their answers before rating the teachers' behavior and they were given as much time as they needed.

While watching the classroom videos, participants were video recorded in a split-screen set-up (recording both the participant as well as the presented classroom videos) for five minutes. This video was used as confrontational video for the think aloud protocols.

3.2.2 Generating Verbal Data

After participants had finished all teacher ratings, the think aloud protocols were generated. Participants watched the split-screen confrontational video of themselves and the classroom video (Fig. 1). We asked participants to think aloud by reporting how their impressions and judgments emerged and how their opinion evolved. In order not to disturb the rating process, we conducted the think aloud protocols afterward. The guided interview consisted of twelve questions (Appendix A), for example: *What was going on in your head while you were watching the videos and completing the rating questionnaires? How did you arrive at your judgment?*

3.3 Stimulusmaterial

The video material was edited from the German IGEL-study (Hardy et al., 2011). The IGEL-study explored third graders knowledge development regarding floating and sinking. In a standardized science lesson, teachers were provided with prepared materials and a script. In the SliceS, a video snippet of ten seconds was randomly edited from each third of the lesson (see Ambady et al., 2000). These three video snippets of ten seconds in length were pasted consecutively for the final thin slice. Hence, the thin slice for each teacher was 30 seconds in length and consisted of three snippets of ten seconds each. However, in all sampled snippets, the teacher was required to be clearly visible. Since Begrich et al.

(2017, 2020, 2021) applied this sampling strategy and demonstrated the accuracy of thin slices ratings of teaching quality, we applied this sampling strategy in the present study as well. Research showed that slices from different phases (beginning, middle, end) of the full footage correlated strongly with each other, indicating good interchangeability of the slices (Hall et al., 2019). In the LongS, a time stamp was generated randomly from which a 10-minute video was edited. All teachers featured in the videos were female.

3.4 Rater Sample and Randomized Assignment

The rater sample consisted of 20 Bachelor and Master students of psychology (two males). Ten raters were randomly assigned to the two groups. Begrich et al. (2017) obtained with a smaller number of thin slices raters (n = 9) robust results. The age of the participants ranged from 18 to 32 years, averaging 24.8 years (SD = 4.2). Participants were recruited via social media. Raters were blind to the specific aim of the study and were unaware of the group they were assigned to or the existence of two groups. All participants received 20€ as a reward.

3.5 Data Collection

The study and data collection were conducted in accordance with the General Data Protection Regulation of the European Union and the German Federal Data Protection Act (BDSG) to which participants agreed with their signature on a letter of consent. According to data protection specifications, the split-screen confrontational videos were deleted under the eye of the participants immediately after the think aloud protocols were generated (see 3.2.1). We utilized various software tools in the process of data collection and data processing: Cisco webex served as the video conference tool, and VLC media player was used for presenting the videos, which were in mp4 format. Participants were video-recorded with 1.8.3.0 screenpresso PRO (2020), and their verbal reports were audio-recorded with 2.4.2 Audacity® recording and editing software (2020). For data analysis, we worked with MAXQDA version 18.2.5. The statistical analyses were conducted in SPSS version 22 and in RStudio version 1.3.1093.

4. Coding and Analysis

Data analysis resulted in three types of information: (1) Codes for text segments of think aloud protocols and guided interviews, (2) the frequency of certain terms used in the verbal material (lexical search), and (3) the amount of verbal data produced by the participants (word count).

4.1 Development of Codes

Codes were developed inductively as well as deductively. Only a few codes were conceived theory-driven before data collection. Most of the codes were developed data-driven (Table 1). A middle-order approach between inductive and deductive coding was applied in the data coding process, meaning we coded the data while having the research aims and relevant theories of social cognition in mind (Saldaña, 2009). From the various terminologies and accounts of dual process theories in the literature, we adopted the System 1 and System 2 terminology (see 1.2). The inferential character of the codes ranged from being high-inferential, requiring the coder to interpret the data, to rather low-inferential, meaning that the coder did not have to abstract or interpret the data. Most of the codes were low-inferential.

All codes and their modes of generation are listed in Table 1 (Classification of Codes) and Table 2 (Codebook). The codes “System 1” and “System 2” refer to processes of System 1 and System 2,

respectively (Evans, 2008; Stanovich et al., 2014). The code "Own School Days" denotes participants' autobiographical memories of their time at school. Due to its associative character, this code serves as an indicator of System 1 (Evans & Stanovich, 2013; Stanovich, 2009). The code "Emotion" is associated with processes of System 1 (Evans, 2008; Kahneman, 2011). The code "Halo effect" is a marker of System 1, indicating evidence for the halo effect (Kahneman, 2011). The code "System 2 overrides System 1" refers to the ability of System 2 to override processes of System 1. The code "Detail" is a marker of processes of System 2 because System 2 allocates attention for a detailed and specific way of processing (Kahneman, 2011). We assume that the underlying cognitive processes of the code "Comparison" include deliberate operations as well as the involvement of working memory (Evans, 2008; Evans & Stanovich, 2013). Further, Kahneman (2011) explains that System 2 is able to compare objects on several aspects. The code "Evaluation" is an indicator of System 2. We consider the underlying cognitive processes of an evaluation to be deliberative, involving the use of working memory (Evans, 2008; Evans & Stanovich, 2013). We consider the two codes "Keep s.th. in mind" and "Item-Situation" as representing cognitive strategies. Both are indicators for System 2 processing (Stanovich, 2009). The three codes "Insufficient Information", "Easiness of Judgment" and "Difficulty of Judgment" are not indicating a specific kind of cognitive processes. However, they provide information about participants' ability to verbalize their cognitive processes while undergoing the thin slices procedure.

Table 1

Classification of Codes

Code	Inference	Cognitive System	Literature
System 1	deductive	System 1	Evans (2008); Stanovich et al. (2014); Wood (2013)
Own School Days	inductive	System 1	Evans and Stanovich (2013); Stanovich (2009)
Emotion	deductive	System 1	Evans (2008); Kahneman (2011)
Halo Effect	deductive	System 1	Kahneman (2011)
System 2	deductive	System 2	Evans (2008); Stanovich et al. (2014); Wood (2013)
System 2 overrides System 1	inductive	System 2	Kahneman (2011); Mugg (2015), Stanovich (2009)
Detail	inductive	System 2	Kahneman (2011)
Comparison	inductive	System 2	Kahneman (2011)
Evaluation	inductive	System 2	Evans (2008); Evans and Stanovich (2013)
Strategy: Keep s.th. in mind	inductive	System 2	Stanovich (2009)
Strategy: Item - Situation	inductive	System 2	Stanovich (2009)
Focus[a]: Teacher	inductive	Not defined	
Focus[a]: Student(s)	inductive	Not defined	
Focus[a]: Situation & Objects	inductive	Not defined	
Insufficient Information	inductive	-	
Difficulty of Judgment	inductive	-	
Easiness of Judgment	inductive	-	

Note. The second column from the left (Inference) denotes whether the code was generated inductively or deductively. In the fourth column (far right), relevant literature is listed. The third column indicates to which cognitive system the code is categorized.

4.2 Coder Training and Intercoder Reliability

Based on the coding scheme, the codebook was developed as a manual for the coders. Cues from the coder training were used to increase distinctness of the codes and comprehensibility of the coding manual. Table 2 provides an excerpt of the codebook. The so-called unitization problem (Campbell et al., 2013; Krippendorff, 2004) was solved by segmenting the data into meaningful conceptual breaks.

Two student assistants were trained as coders in 12 sessions over six weeks (Goodell et al., 2016). Audio records from the trial were transcribed for the training of the coders. During the process of training, transcripts from the trial were scored independently by the two coders and subsequently compared and discussed. The two coders scored the data blindly in terms of not knowing to what rater group a transcript belonged. All training sessions took place via video call.

To calculate the intercoder reliability, three out of 20 transcripts (15% of the data) from the main study were independently double-scored by both coders, as recommended by O'Connor and Joffe (2020). With kappa of .74 over all three transcripts (.71, .73., and .76 for the first, second, and third transcript, respectively), a good intercoder reliability was achieved (Brennan & Prediger, 1981; Rädiker & Kuckartz, 2019). Subsequently, the two coders reached a discursive agreement on the final versions of the three transcripts they had initially scored independently. Finally, one coder scored the remaining 17 transcripts. The scoring process was supervised, discussed and guided by the first author in order to ensure adherence to quality standards.

Table 2

Codebook

Code	Definition/Description	Examples
System 1	Statements of participants reflecting System 1 of dual system accounts of social cognition. Impression formation is fast, intuitive, associative, emotionally, automatic and unconscious (Evans, 2008; Stanovich et al., 2014; Wood, 2013).	„Deciding was hard because everything happened so quick that I could not think about it. I had to decide intuitively guided by my emotions." "The first impression was by the guts. It was the feeling the teacher gave me."
Own School Days	Participants talk about autobiographic memories and experiences of their own school days back in the days when they were students. They mention their own teachers, instructions or classes.	„… when I was a student…" "During my school days…" "My teacher in elementary school…" "My instruction in elementary school…"
Emotion	Participants report feelings, emotions or affects triggered by the video. These emotions may occur during the video or during the questionnaire. The emotion needs to be self-referenced. This does not include emotions ascribed to the students (e.g. "The students are laughing") or the teacher.	„It was very exciting watching these kids…" "The pupils were so cute…" "It was nice seeing that…"
Halo-Effect	A positive or negative aspect of the teacher influences other aspects. This also incorporates sympathy or antipathy. We also include remarks of participants when they are trying to avoid the Halo-Effect.	„The permanent admonishments of the teacher overshadowed pretty much." Although the teacher was pretty unappealing, I tried to judge him fairly."
System 2	Statements of participants reflecting System 2 of dual system accounts of social cognition. This refers to i.a. analytic, reflective, rational, deliberate, effortful, complex thinking (Evans, 2008; Stanovich et al., 2014; Wood, 2013).	"I was thinking about my decision, and my judgments were based on a pattern." "I was thinking about it and analyzed it."
System 2 overrides System 1	At first System 1 is active and then over time System 2 comes into play. This is indicated for instance when 1.) participants state that they changed the tick in the questionnaire, 2.) when participants mention a gain in knowledge, or 3.) when participants changed their mind/opinion.	„At first, I went by my gut but then the longer the video was I've found more and more examples, which made me change my mind." "At first, I thought that the teacher was bored and not motivated, but then when I was thinking about it,

		I realized that the teacher was not that bad, and I changed my mind." "At first, I went by the guts, but then my deliberate thinking was also involved."
Detail	A detail is a precisely contoured unity within a larger context. This includes for example: • a concrete action/behavior (e.g. smile, gesture, statements/quotes) in a specific moment, • clothes, haircuts, pictures, writings on the blackboard, • participants mentioning that they were able to name details, • names of students	„Through the window, I saw an apartment block." "In fact, I was looking out for details." "The teacher gave a student a yellow card because he was talking." "I think, it was very rude that the teacher made *ssshhh*, when the boy laughed loudly." "There was this moment when a girl turned around and…"
Comparison	Participants compare (aspects of) teachers or videos. This includes the comparative and superlative. Participants use a video as anchor in order to compare.	„I always compared the teachers. I realized that the second teacher was in in comparison to the first teacher much more… "
Evaluation	An evaluation can be situated on a continuum from good to bad and refers to actions, persons or situations. Hence, evaluations could be ranked in an order. An evaluation contains critique or praise and would potentially evoke an emotion or affect in the evaluated person. Evaluative adjectives are included.	"The teacher didn't do it so well…" "She wasn't a good teacher." "She seemed to be very a very nice person." "The whole group did work very well together and the atmosphere was positive."
Strategy Keep s.th. in mind	Participants trying to keep items of the questionnaire in mind while watching the video. They use the information of the questionnaire as guide while watching the video.	„I tried to memorize the items and watched so the videos." "I kept the items in mind so I knew whereon I had to look for."
Strategy Item-Situation	1. While completing the questionnaire participants trying actively to think back to the video. 2. While watching the video participants actively tried to memorize aspects or situations of the video in order to answer the questionnaire.	„I watched the video and thought what could be helpful for answering the questionnaire." "I tried to memorize issues of the video which could be helpful in order to answer the questions"
Focus[a] Teacher	Participants focus (commenting/talking) on the teacher. Participants trying to infer from hints of teachers' behavior to answer the questionnaire. This did also include reports about nonverbal cues like gesture, posture, facial expression, tone of voice.	„The teacher interrupted one student that's why I answered in the questionnaire …"
Focus[a] Student(s)	Participants focus (commenting/talking) on a/the student(s). Participants trying to infer from hints of students' behavior to answer the questionnaire.	„Many students put up their hands. That's why I thought that the teacher is able to involve all pupils."
Focus[a] Situation & Objects	Participants focus (commenting/talking) on the situation or objects. Participants trying to infer from hints of the situation or objects to answer the questionnaire. This includes remarks about the atmosphere.	„The posters on the wall seemed to be very tidy. That is why I thought that …"

Insufficient Information	Participants claimed that 1. the video (field of view) did not contain relevant information about specific traits/items, 2.the video was too short. The participants mention that they would have liked to see more. They state that they did not know something or could not detect anything within the context of the video.	„In the video, I could not see whether the teacher was nice." The video was so short that I could not figure out was it was all about." "No concrete situation was displayed about the first items…"
Difficulty or Easiness of Judgments	Participants mention that judgments or items were difficult or easy to answer. We also included complaints about the Likert scale being too narrow.	"Answering the first item was pretty hard." "Judging teachers based only on so extreme short videos was very hard." "This is a tough question." "Some items were quite easy to answer."

[a]Focus of attention

4.3 Descriptive Analysis

Based on Namey et al. (2008), code frequency lists were generated by counting the number of codes (Table 3). The code frequency lists were developed with two different approaches: counting all codes[1] (sum of codes across all individuals for each rater group) and counting all individuals[2] (sum of individuals a code was ascribed to at least once for each rater group). To compare group means between the two rater groups, we calculated t-Tests or Welch's t-Tests for the sum of codes across all individuals[2] or the Exact Test of Fisher for the sum of individuals[3] a code was ascribed to at least once (Namey et al., 2008). Moreover, for the sum of codes, the minimum of applications of a code for a person and the maximum of applications of a code for a person as well as the standard deviations are presented in Table 3[3].

[1] Table 3 Sum of codes: absolute (third column)
[2] Table 3 Sum of individuals (column far right)
[3] Table 3 Sum of codes: min.-max. (Standard deviation) (fourth column)

Table 3

Results of the Frequency Counts

Code	Condition	[a]Sum of codes: absolute	Sum of Codes: min.-max. (Standard deviation)	[b]Corrected sum of codes: Relation to words	[c]Sum of individuals
System 1	SliceS	25	1-6 (1.72)	0.19	10
	LongS	21	1-6 (1.73)	0.11	10
Own School Days	SliceS	13	0-3 (1.06)	0.10	8
	LongS	2*	0-1 (0.44)	0.01	2*
Emotion	SliceS	31	1-7 (2.03)	0.23	10
	LongS	64	0-16 (5.95)	0.32	9
Halo-Effect	SliceS	26	0-6 (1.65)	0.20	9
	LongS	19	0-5 (1.52)	0.10	9
System 2	SliceS	5	0-3 (0.97)	0.04	3
	LongS	22**	0-5 (1.48)	0.11	9*
System 2 overrides System 1	SliceS	4	0-3 (0.97)	0.03	2
	LongS	19**	0-3 (0.99)	0.10	9**
Detail	SliceS	7	0-4 (1.34)	0.05	3
	LongS	39**	1-10 (3.49)	0.20	10**
Comparison	SliceS	8	0-2 (0.79)	0.06	6
	LongS	40*	2-9 (2.16)	0.20	10
Evaluation	SliceS	69	3-11 (2.73)	0.52	10
	LongS	148**	6-28 (7.33)	0.75	10
Keep s.th. in mind	SliceS	9	0-4 (1.45)	0.07	4
	LongS	12	0-3 (1.23)	0.06	6
Item-Situation	SliceS	5	0-3 (0.97)	0.04	3
	LongS	13	0-3 (1.16)	0.07	7
Teacher	SliceS	36	1-7 (1.90)	0.27	9
	LongS	31	1-5 (1.45)	0.16	9
Student	SliceS	13	0-3 (0.95)	0.10	8
	LongS	15	0-5 (1.43)	0.08	8
Situation & Objects	SliceS	9	0-3 (0.99)	0.07	6
	LongS	7	0-2 (0.68)	0.04	6
Insufficient Information	SliceS	30	0-8 (2.63)	0.23	9
	LongS	6*	0-4 (1.27)	0.03	3*
Difficulty of Judgment	SliceS	35	1-7 (1.90)	0.26	10
	LongS	29	1-6 (1.45)	0.15	10
Easiness of Judgment	SliceS	8	0-2 (0.63)	0.06	7
	LongS	8	0-2 (0.79)	0.04	6

*Test comparing the two rater groups is significant at the 0.05 level (two-tailed).

**Test comparing the two rater groups is significant at the 0.01 level (two-tailed).

[a] t-Test or Welch's t-Test (sum of codes across all individuals for each rater group and standard deviation)

[b] Number of codes divided by number of words in total and multiplied by 100

[c] Exact Test of Fisher (sum of individuals a code was ascribed to at least once for each rater group)

4.4 Lexical Search

A lexical search was conducted, using the built-in lexical search function in MAXQDA. Signal words or lexical phrases were categorized as either referring to processes of System 1 or processes of System 2 (Table 4). Based on the literature, the terms "quick/fast", "intuitive" and "automatic" signal System 1 processing (Evans, 2008; Wood, 2014), whereas "criterion" and "reflect(ed)" are associated with processes of System 2 (Evans, 2008; Stanovich, 2009). Additionally, more signal words were generated data-driven in an exploratory mode. The word "warmth" was interpreted as an indicator of emotional processing, denoting System 1. The word "atmosphere" is considered a holistic approach to information processing, indicating System 1. The three search terms "fidget/fidgety", "noisy" and "disturbances" could not clearly be allocated to one of the two cognitive system. For each rater group, the number of signal words or lexical phrases was counted, and the relevant findings are listed in Table 4.

Table 4

Lexical Search

Keyword	Thin Slices Situation	Long Video Situation	Cognitive System
automatic	0	4	1
intuitive	2	1	1
quick/fast	16	5	1
atmosphere	14	3	1
warmth	4	1	1
criterion	3	0	2
reflect(-ed)	3	4	2
fidget/fidgety	11	3	-
noisy	19	0	-
disturbances	22	20	-

Note. Absolute number of appearances for each condition.

4.5 Word Count and Relative Occurrence of Codes

An overall word count of participants' answers was conducted to examine in which rater group more words were uttered—whether participants in the SliceS uttered more (or fewer) words than participants in the LongS. Hence, the number of words for each participant in both rating situation was summed up. In addition, we wanted to prevent any distorting effects due to possible differences in the amount of words per answer. Therefore, the frequency of codes was related to the total number of words for both rater groups. We divided the number of codes in total by the number of words in total for both rating situations and multiplied it by 100 (Table 3)[4]. All irrelevant communication content, such as interview questions, researcher comments, and off-topic remarks, were removed from the data, leaving only the participants' responses for analysis.

4.6 Further Analysis

For some codes, the scored text segments were analyzed in depth to uncover further hidden underlying patterns in the data. We did this for the code "Evaluation" by examining whether the evaluations were positive, neutral, or negative. Further, we analyzed the text segments of the code "System 2 overrides System 1" of participants in the LongS. We examined whether an initial first impression changed in the course of judgment formation entirely or only slightly. Moreover, we also examined whether a change in participants' judgments was abrupt or whether the change evolved gradually. Finally, some codes were deleted due to the scarcity of their application (Cope, 2010; Miles & Huberman, 1994).

5. Results

The current study explored the cognitive mechanisms underlying thin slices ratings of instructional quality. Since no previous study has examined this phenomenon, we employed both inductive and deductive analysis approaches. The verbal data were analyzed qualitatively, resulting in a set of codes that indicated cognitive mechanisms in relation to impression formation. The codes were then statistically analyzed to compare the two rating situations.

[4] Table 3 Corrected sum of codes: Relation to words (Fifth column)

5.1 Preliminary Analysis

In preliminary analyses, we tested whether our overall design worked by examining whether raters were able to report about their judgment formation processes. All participants in the thin slices rating situation (SliceS) could answer all questions. In total, the verbal data of the ten participants in the SliceS consisted of 13.277 words, with an average of 1.328 words per participant (*SD* = 419 words). In comparison, participants in the Long Video Situation (LongS) answered with a total of 19.773 words, with an average of 1.977 words per participant (*SD* = 937 words). The code "Easiness of Judgment" revealed no substantial differences between the groups. The code "Difficulty of Judgment" showed that participants in the SliceS mentioned somewhat more often (0.26 vs. 0.15) difficulties than participants in the LongS. However, all ten participants in the LongS did this to a certain extent as well (Table 3). Moreover, significantly more participants in the SliceS (9 vs. 3) complained significantly more often (30 vs. 6) about the scarcity of information than participants in the LongS, which was indicated by the code "Insufficient Information" (Table 3).

5.2 System 1 Processing as the Foundation of Thin Slices Ratings (Research Question 1)

To address research question 1, we examined the verbal data for evidence of typical System 1 processes, expecting to find more evidence for System 1 processes in the verbal reports of participants undergoing the SliceS. Descriptive results are presented in Table 3. With respect to the sum of codes (Table 3; third column), all seven significant test results indicated group differences as expected in our research questions. Concerning the sum of individuals (Table 3; column far right), all five significant test results indicated differences between the two groups as expected.

The direct coding of supposed System 1 processes is indicated with the code "System 1". In relation to words, evidence of System 1 was found more often in the SliceS (0.19) than in the LongS (0.11), meaning that per 100 words the code "System 1" was applied 0.19 in the SliceS and only 0.11 times in the LongS (Table 3). However, the difference between the conditions was not statistically significant. In the verbal data of all 20 participants, both, in SliceS as well as in the LongS, direct evidence for System 1 was found, meaning that all participants operated in System 1 at some stage. The finding that thin slices raters operated in System 1, as did participants in the LongS, is predominantly in accordance with our hypothesis.

The code "Own School Days" is an indicator of System 1. The code was more prevalent in the SliceS (Table 3). Not only did participants in the SliceS refer more often to their own school biography (13 vs. 2), but also more participants did so in the SliceS than in the LongS (8 vs. 2; difference tests were significant for "sum of codes" and "sum of participants"). This result is in line with our expectations.

The code "Emotion" is an indicator for processes of System 1. All 10 participants in the SliceS and nine participants in the LongS reported at least once an emotion. However, in relation to words, participants in the SliceS reported fewer emotions (0.23) than participants in the LongS (0.32), yet the difference was not statistically significant (Table 3). Further, we searched all text segments, which were scored with the code "Emotion" whether the expressed emotions were positive or negative. Yet, no interesting results distinguishing the two rater groups emerged. Summarizing, participants in both rater groups reported an emotion similarly often. This result is not in line with our expectations.

The code "Halo Effect" is an indicator for processes of System 1. In the comments of nine participants in both conditions evidence for the halo effect was found. In relation to words, more evidence for the halo effect was found in the SliceS (0.20) than in the LongS (0.10), but the difference was not statistically significant (Table 3). Summarizing the results, the evidence suggest that thin slices raters are prone to the halo effect—but not exclusively. Evidence for the halo effect was found slightly more often, though not significantly, in the verbal reports of participants in the SliceS. This means that this finding rather supports our research expectations.

We assume that the code "Evaluation" is associated with processes of System 2 (see also 5.3). We analyzed whether participants' evaluations were positive, neutral, or negative by conducting a one-

way ANOVA. Evaluations in the SliceS were negative 37 times (M = 3.7, SD = 1.70), neutral 14 times (M = 1.4, SD = 1.17), and positive 19 times (M = 1.9, SD = 1.52), whereby the differences between negative, neutral, and positive evaluations were statistically significant [F(2, 27) = 6.65, p = .004]. In the LongS, exactly the opposite occurred: More evaluations were positive (86 times) than negative (45 times) or neutral (25 times). Likewise, the difference was statistically significant [F(2, 27) = 9.02, p = .001]. In sum, the evidence suggests that thin slices raters' evaluations were significantly more often negative than positive.

The three codes "Teacher", "Student(s)" and "Situation & Objects", indicating participants' focus of attention, are not associated with System 1 or System 2 functioning (Table 1). No significant results were found distinguishing the two rater groups (Table 3). Consequently, these codes are not further discussed.

The lexical search revealed that the terms "atmosphere" and "quick/fast", which we consider as indicators of System 1 occurred considerably more frequently in the SliceS in comparison to the LongS (Table 4), even though participants in the SliceS produced on average fewer words than participants in the LongS. These results support our claim that processes of System 1 are the foundation of thin slices ratings.

5.3 Are Processes of System 2 dissimilar to Cognitive Processes of Thin Slices Ratings (Research Question 2)?

In order to analyze whether the underlying cognitive processes of thin slices ratings are sufficiently dissimilar to those of System 2, we examined the participants' verbal data for evidence of typical processes of System 2 and whether such evidence occurred more frequently in the LongS.

The direct code "System 2" refers to processes of System 2 (Table 2). Significantly less participants in the SliceS (3 vs. 9) operated significantly less often in System 2 (5 vs. 22) than participants in the LongS (Table 3). Moreover, in relation to words, participants operated less often in System 2 in the SliceS (0.04) in comparison to the LongS (0.11; Table 3). In sum, these findings support our research expectation.

The code "System 2 overrides System 1" is associated with cognitive operations of System 2 (Table 2). Significantly less participants in the SliceS (2 vs. 9) changed their impression significantly less often (4 vs. 19 times) than participants in the LongS (Table 3). Further, taken the relative occurrence (Table 3) into account, it seems that participants in the SliceS (0.03) changed a first impression less often than participants in the LongS (0.10). We analyzed the verbal data regarding the code "System 2 overrides System 1" only for the LongS in depth (see section 4.5) in order to analyze whether an initial first impression changed in the course of impression formation entirely or only slightly. We found that only once a participant changed his/her impression about the teacher entirely, whereas eight times an impression was corrected or refined only slightly. Moreover, we wanted to determine whether the change of the initial first impression was sudden and abrupt or whether it gradually evolved. We found no evidence in the data that the change was abrupt or sudden, but in six cases the final judgment evolved gradually. Summarizing the results, in contrast to raters relying on more information (i.e., 10-min videos), it seems that thin slices raters of instructional quality only rarely change an initial first impression.

The code "Detail" can be considered as an indicator of System 2 processing. In the SliceS, significantly less participants (3 vs. 10) remembered significantly less details (7 vs. 39) than participants in the LongS. Controlling for possible distorting effects due the length of the answers, participants reports were less detailed in the SliceS (0.05 vs. 0.20; Table 3). In sum, it seems that reports of participants in the SliceS are less detailed, which aligns with our research expectation.

The code "Comparison" is an indicator of System 2. In the SliceS, less participants (6 vs. 10) compared significantly less often (8 vs. 40) the given information than participants in the LongS (Table 3). Correcting for the number of words, participants in the SliceS compared the information less often

(0.06 vs. 0.20) than in the LongS. In sum, it seems that participants in the SliceS less frequently engage in comparing information, which we consider as supporting evidence for our research question.

The code "Evaluation" is an indicator of System 2. In the SliceS as well as in the LongS every single participant evaluated the given information in some way. In total, participants in the SliceS evaluated the information significantly less often (69 vs. 148) than participants in the LongS. Correcting for the number of words, participants in the SliceS evaluated the information less often (0.52 vs. 0.75). In line with the research expectation, participants in the SliceS evaluate the given information less often.

The two codes "Keep s.th. in mind" and "Item-Situation" are representing cognitive strategies and both are indicators for System 2 processing. No statistically significant differences were found between the groups (Table 3). However, slightly more evidence for cognitive strategies were found in participants' data in the LongS. In sum, we claim to have found some evidence for the presence of cognitive strategies in judgment formation of thin slices raters. This result is not quite in accordance with our research expectation.

6. Discussion

Assessing instructional quality in schools and in ECEC is very costly and labor-intensive (Murphy & Hall, 2021). Therefore, a more economical and yet accurate approach would be desirable. Previous research has demonstrated that raters, relying on minimal information, can accurately assess teaching quality in schools and ECEC (Ambady & Rosenthal, 1993; Begrich et al., 2021; Sokolovic et al., 2021; Vinokic et al., 2024). The present study examined the underlying cognitive processes of ratings based on first impressions of instructional quality (i.e., thin slices ratings). To our knowledge, no empirical evidence has been documented that directly addresses this issue. Based on the literature of dual process theory, we expected to find evidence of typical System 1 processes and little to no evidence of System 2 processing in the verbal data of thin slices raters. Gaining insights into the cognitive mechanisms shaping impression formation in thin slices raters may help to optimize the practical application of the thin slices technique as a trustful measurement method to complement the established repertoire of measurement techniques in the field of teaching quality.

In the following section, the discussion focuses on embedding this study's results in the current literature on dual process theory, classroom observation methodology and thin slices research. Initially, we will discuss whether participants are able to retrospectively verbalize their cognitive processes. Subsequently, we will discuss whether typical processes of System 1 are the foundation of thin slices ratings (RQ1) and whether thin slices ratings are dissimilar to typical processes of System 2 (RQ2).

Are participants in thin slices rating situations of instructional quality able to report about their judgmental processes at all? Only a few studies have examined the level of awareness associated with first impressions, whereby the evidence is mixed (Ames et al., 2010; Biesanz et al., 2011). Given that participants are able to report about their judgmental processes, are the verbal reports expedient and useful for drawing conclusions about the underlying cognitive processes of thin slices raters? Various pieces of evidence suggest that thin slices raters are indeed able to report about their judgmental processes. Both thin slices raters as well as raters relying on more information equally complained about the difficulty of the tasks, indicating that the two different rating situations did not seem to influence the perceived difficulty of assessing instructional quality or verbalizing mental processes. The answers of thin slices raters differed significantly in many aspects from the answers of participants relying on more information (i.e., 10 minutes). The fact that statistically significant results were found between the two rater groups may be interpreted as a hint of different cognitive processes at work. The presence of clear and statistically significant patterns distinguishing the two rater groups supports the idea of different processes of social perception. Moreover, the significant patterns distinguishing the two groups are in line with the current body of literature about dual process theories. This allows the conclusion that thin slices raters are indeed able to report about their underlying cognitive processes, and that these processes are indeed different from participants in the Long Video Situation. Or, to put it differently: If System 1

functioning would not manifest in the verbal data, our analysis probably would not have detected statistically significant patterns.

6.1 System 1 Processing as the Foundation of Thin Slices Ratings

All participants in both groups seem to have relied equally on System 1. However, thin slices judgments rely to a lesser degree on System 2 functioning compared to judgments based on more information (i.e., 10-minutes videos; LongS). The following evidence suggests that processes of Systems 1 are the foundation of thin slices ratings.

A defining feature of System 1 processes is autonomy. The execution of autonomous processes tends to be associative (Evans & Stanovich, 2013; Sloman, 1996). Thin slices raters more frequently referred to biographic information in terms of associations with their earlier school life (e.g., their elementary school teachers). In comparison, participants exposed to more information had far fewer associations with their own school days. Associations are considered a typical feature of System 1 processing. This supports the assumption that System 1 is active while undergoing a typical thin slices rating situation. The fact that these associations are related to teaching or school is not trivial since it may have consequences for rating results received via the thin slices technique. Positive or negative biographic memories respectively individual experiences may impact thin slices ratings of instructional quality. Further investigations of a potential effect of individual experiences on the accuracy of thin slices ratings are necessary.

As expected, several of the codes that reflected System 1 processing occurred particularly often in the Thin Slices Situation, yet some unexpected results emerged. We expected more emotions to occur in judgments based on first impressions because emotional processing is explicitly linked to System 1 (Evans, 2008; Kahneman, 2011). However, we found the opposite: more emotions were detected in judgments based on more information instead of in judgments based on first impressions. A particularity of the German language might contribute at least partly to this finding, though. The German word for feeling *("Gefühl")* can carry two semantically different meanings: One refers to an emotion or a feeling and the other refers to an assumption or a hunch. Hence, the validity of this code may be doubted.

The halo effect is a type of cognitive bias in which one aspect of a person influences the judgment of other aspects—a bias that is seen as typical for System 1 (e.g., Kahneman, 2011). In our study, the majority of thin slices raters indeed exhibited the halo effect, indicating a superficial, holistic cognitive approach. However, raters exposed to more information were also prone to the halo effect, albeit to a somewhat lesser degree, suggesting that they, too, possibly processed information in System 1. In sum, it seems that thin slices raters of instructional quality are prone to the halo effect, but not notably more prone compared to judgments based on more information.

All thin slices raters evaluated the information given in the video. From an evolutionary perspective, first impressions tend to be easily pulled to the negative end of the evaluative dimension. The fitness costs of an incorrect negative first impression are potentially lower than the fitness costs of an incorrect positive first impression (Ambady & Skowronski, 2008; Nesse, 2005). A strong negative interpersonal impression may be formed on the basis of very little information, and a positive interpersonal impression may require a greater amount of informational input (Ambady & Skowronski, 2008). The fact that significantly more evaluations were negative than positive in verbal reports of thin slices raters is in line with Ambady and Skowronski (2008), as well as the fact that we found significantly more positive evaluations than negative in judgments based on more information. In sum, we claim to have found robust evidence that thin slices raters' evaluations are rather negatively than positively.

The results of the lexical search indicated that participants assessing instructional quality based on their first impression might operate holistically. Nisbett et al. (2001) define holistic thought as involving an orientation to the context or the field as a whole, claiming that System 1 operates holistically. The word "atmosphere" (Table 4) was used more frequently by thin slices raters than by

raters relying on more information. In the context of assessing classroom videos, we interpret the meaning of the word "atmosphere" as referring to the situation globally and as a whole, implying that "atmosphere" is a marker of System 1. In sum, we claim to have found some evidence for holistic information processing in thin slices judgments.

In summary, assessing instructional quality solely based on first impressions (i.e., thin slices ratings) appears to be associatively, rather negatively than positively, prone to the halo effect, and probably holistically. According to the literature of dual process theories of social cognition, these results can be considered as typical processes of System 1. Accordingly, System 1 appears to be the underlying cognitive system of thin slices ratings of instructional quality. Therefore, research question 1 can be considered confirmed.

6.2 Are Processes of System 2 dissimilar to Cognitive Processes of Thin Slices Ratings?

All participants appeared to rely equally on System 1. However, thin slices judgments seem to rely to a much lesser degree on System 2 functioning than judgments based on more information (i.e., 10-minute videos), which seem to involve both systems—System 1 as well as System 2. The data lead to the conclusion that social cognition is initially dominated by System 1 processes, with System 2 being activated after a while. It appears that System 1 processes constitute the essential first phase of social cognition. These findings align with Mugg (2015) who claims that Type-1 and Type-2 processes are generated at different times: At first Type-1 is activated and then, under some circumstances, Type-2 processes start to work. Type-2 processing occasionally overrides or intervenes on Type-1 responses rather than producing responses all on its own (Mugg, 2015). Moreover, our results suggest that raters with access to more information (i.e., 10-minute videos) altered their initial impression much more frequently than thin slices raters. We found no evidence in the data that this change was abrupt or sudden. Instead, we found various hints of evidence that the judgment evolved gradually (see 5.1). Based on the results, we may claim that the transition from System 1 processing to System 2 processing is not abrupt; rather, we can understand it as a gradual, with System 2 progressively taking over. Our results do not provide an explanation of *when* System 2 is activated; however, according to our design, it appears to occur around or after 30 seconds, but within ten minutes. Furthermore, we discovered evidence that System 2 processing did not entirely nullify or override System 1 judgments but rather shaped, modified or refined them. Only one participant mentioned once that she/he entirely changed the initial first impression after a while. However, considerably more evidence was detected that the initial first impression was only slightly or to a minor extent modified. Therefore, in hindsight, a more suitable name for the code would have been "System 2 modifies System 1". In sum, we claim to have found evidence that System 2 is not the foundation of thin slices ratings, or only to a substantially lesser degree compared to judgments based on more information.

Kahneman (2011) posits that more detailed and specific processing of information is a feature of System 2. Deeper cognitive processing should enhance the encoding and recall of details. Considering that thin slices judgments do not rely on deep elaboration (i.e., System 2), thin slices raters should report less details than participants exposed to more information (10-minute videos). Our study confirms this assumption. However, the research design may have confounding effects on the results. The amount of information provided, such as the length of the video, rather than the depth of processing, may have caused this finding. It is possible that longer videos simply lead to more details being memorized and subsequently recalled.

We scrutinized the body of literature to determine which cognitive system generates comparisons. System 2 can compare objects based on several attributes (Kahneman, 2011). Participants exposed to more information made considerably more comparisons than thin slices raters. Consequently, this finding appears to be evidence for the dissimilarity of typical processes of System 2 and the cognitive processes underlying thin slices ratings.

Based on what we know about System 2 functioning, we presumed that an evaluation can be considered as a deliberate act involving the working memory and can therefore be subsumed to the processes of System 2 (Evans, 2008; Evans & Stanovich et al., 2014). Given that System 2 processes are not at activate in thin slices judgments, we did not expect thin slices raters to evaluate the given information. Drawing on the evidence, thin slices raters evaluated the information to a lesser extent compared to raters with access to more information, highlighting a clear distinction between cognitive processes underlying thin slices ratings and processes of System 2.

In the present study, a cognitive strategy was operationalized in a broader sense. Two codes, "Keep s.th. in mind" and "Item-Situation" (Table 1 and 2), represent cognitive strategies. They were generated inductively based on the data and are very narrowly defined in comparison to what is considered a cognitive strategy in the literature. However, they summarized how participants proceeded and how they tried to solve the problem. The results indicate that the two codes occurred slightly more often in judgments based on more information than in judgments based on first impressions. The code "Keep s.th. in mind" refers to the strategy of memorizing the rating items before watching the next classroom video. Hence, it is not clear whether this code is truly a valid indicator for a cognitive strategy because some cognitive processes occurred already before the rating situation started (Fig. 1). The code "Item-Situation" refers to participants trying to actively retrieving contents of the video while filling in the rating questionnaire. Less thin slices raters tried to actively recall content from the video while filling in the rating questionnaire in contrast to participants observing 10-minute videos. In sum, these findings are partly in line with Stanovich (2009) because he exclusively attributes strategic operation as being part of System 2 processes.

In summary, we claim that research question 2 is confirmed, as the cognitive processes underlying thin slices ratings of teaching quality seem to be dissimilar from typical processes of System 2. Based on our data, we posit that thin slices judgments derive very little from analytic or reflective processes. Instead, thin slices judgments predominantly rely on System 1 processes. In comparison, the judgments of participants assessing teaching quality based on 10-minute video clips appear to be initially dominated by System 1 processes, but over time, these judgments are gradually refined and modified with the involvement of System 2.

7. Limitations and Future Directions

The present studies pursued a rather innovative and unconventional approach. To obtain some humble glimpses into the black box of the human mind, we set up two rating situations by implementing a typical thin slices situation and by approximating the rating procedure of a typical classroom observation study (Hardy et al., 2011; DESI-consortium, 2008). The aim was to detect the cognitive processes underlying thin slices ratings of teaching quality. Although the results seem to be in line with the body of literature, some limitations need to be pointed out.

In total, the rater sample consisted of 20 psychology undergraduates. While the sample size was relatively small, it was deemed adequate for our study. (see Begrich et al., 2017). However, further research with a larger sample could carve out some cognitive processes or other phenomena more precisely, resulting in a refinement of the codes. Moreover, a larger sample could contribute to a more robust interpretation and generalization of the results. In particular conducting the study with a different rater sample would be interesting. For instance, thin slices raters mentioned frequently associations to their own school days. These associations had the same content (e.g., school, teachers, classes) as the stimulus material (classroom videos). Only undergraduate psychology students were invited as raters in the present study. Conceivably, this rater population has rather positive associations to their own schooling. A different rater population, with rather negative associations to their own schooling, would potentially produce different results. Further research should consist of a more diverse sample of raters. In a current study, we examine the accuracy of diverse thin slices rater samples with varying levels of expertise (children, undergraduates, teacher trainees, educational experts and adults with unfortunate educational careers).

In retrospective think aloud protocols, verbalization problems might arise, whereas concurrent think aloud protocols might negatively impact task performance (Van Den Haak, 2003). To avoid interference with the rating process, the think aloud protocols were generated retrospectively. Therefore, we used the confrontational video as a stimulation for participants. In both rating situations (SliceS and LongS), retrospective think aloud protocols were used alike. The results should not be influenced by the retrospective think aloud protocols because they were applied in both rater groups.

In order to explore the cognitive processes underlying thin slices ratings, we decided to develop an innovative, unconventional and risk-taking approach. To obtain meaningful results, we designed a typical thin slices rating situation and intended to approximate a conventional observer rating situation (e.g., Hardy et al., 2011; DESI-consortium, 2008). Various parameters of the research design could have been varied, or even a completely different design could have been set up. In the thin slices rating situation, the time for completing the questionnaire was restricted to 30 seconds. Young, educated adults, who had time to study the items before the experiment started, can answer six simple items, recurring ten times for each teacher, in 30 seconds. Our previous thin slices research indicated that thin slices raters do not need more than 30 seconds to complete the rating of six simple items. Further, this is sustained by the very low rate of missing data in the rating items (0.7 %). Moreover, inferring from survey research (Yan & Tourangeau, 2008), we conclude that the answering time for six items should not exceed 30 seconds.

8. Theoretical and Practical Implication

Untrained raters without teaching experience or didactic knowledge are able to assess instructional quality reliably and validly solely on the basis of 30-seconds classroom videos (Begrich et al., 2017, 2020, 2021). Researcher applying the thin slices technique to assess instructional quality often encounter skepticism and disbelief from fellow colleagues from the scientific community concerning the accuracy of thin slices ratings. How is such accuracy possible without analyzing and reflecting properly? The present study may contribute a piece to the puzzle on how thin slices ratings can yield accurate results without actively analyzing longer classroom videos (e.g., 10 minutes, 45 minutes or even longer). The fact that thin slices raters cannot actively analyze, reflect or think about the classroom videos is perhaps the clue to the solution. Thin slices raters rely on a different, highly powerful cognitive system of information processing: System 1 of dual process theories of social cognition (Kahneman, 2011; Stanovich et al. 2014). Conventional raters assessing full-length classroom videos rely predominantly on System 2 (and probably to a minor extent on System 1). Strictly speaking, the fact that thin slices raters do not have enough time to carefully analyze the given information is actually not a drawback but a benefit because different cognitive processes are dominating judgment formation in comparison to conventional ratings. This implies that thin slices ratings could become a promising alternative to other forms of ratings.

Our study provides novel impulses for two different research strands: (a) for those who want to use or have already used the thin slices technique, the results provide insights into the cognitive processes involved in judgment formation and potential biases, and (b) for those working within the conventional paradigm of systematic video ratings of instructional processes. In the present study, the judgments of participants in a rating situation approximating a conventional observer study (10-minute videos) also appear to be influenced by System 1. What conclusion can be drawn from this finding for conventional observer studies (e.g., raters assessing instructional quality based on full-length videos) and their training programs? Does the involvement of System 1 in conventional rating situations play an important yet underestimated role in judgment formation? Can raters be trained to be vigilant about the influence of an initial first impression? Should conventional raters be instructed to refrain from letting their judgments be guided by their first impressions? Could this result in a diminished influence of System 1 processes, causing raters to operate predominantly in System 2? However, it is questionable whether processes of System 1 can actually be suppressed deliberately. Does the intentional attempt to exclude or to reduce the influence of System 1 processes in conventional ratings of teaching quality worsen the accuracy of these ratings. Further research is needed to unravel the interplay and distinct

influences of System 1 and System 2 on impression formation in conventional rating situations of instructional quality.

Keypoints

- Thin slices ratings of teaching quality, based on 30-second classroom videos, appear to rely on typical cognitive processes of System 1 of dual process theories of social cognition.
- Thin slices ratings of teaching quality are associative and tend to be rather negative than positive.
- Ratings of teaching quality based on 10-minute classroom videos rely on both System 1 and System 2 of dual process theories of social cognition, with System 2 possibly modifying an initial judgment.

References

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153–156. https://doi.org/10.1023/A:1008168421283

Ambady, N. (2010). The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry*, *21*(4), 271–278. https://doi.org/ 10.1037/0022-3514.64.3.431

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, *32*, 201–271. https://doi.org/10.1016/S0065-2601(00)80006-4

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431–441. https://doi.org/10.1037/0022-3514.64.3.431

Ambady, N., & Skowronski, J. J. (Eds.). (2008). *First impressions*. Guilford Press.

Ames, D. R., Kammrath, L. K., Suppes, A., & Bolger, N. (2010). Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin-slice impressions. *Personality and Social Psychology Bulletin*, *36*(2), 264–277. https://doi.org/10.1177/0146167209354519

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Audacity Team. (2020). *Audacity recording and editing software* (Version 2.4.2) [Computer software]. https://www.audacity.de/

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. https://doi.org/10.1007/s10919-005-2744-y

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

Begrich, L., Fauth, B., & Kunter, M. (2020). Who sees the most? Differences in students' and educational research experts' first impressions of classroom instruction. *Social Psychology of Education*, *23*(3), 673–699. https://doi.org/10.1007/s11218-020-09554-2

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

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

Bellini-Leite, S. C. (2018). Dual process theory: Systems, types, minds, modes, kinds or metaphors? A critical review. *Review of Philosophy and Psychology*, *9*(2), 213–225. https://doi.org/10.1007/s13164-017-0376-x

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, *56*(3), 205–213. https://doi.org/10.1177/0022487105275904

Biesanz, J. C., Human, L. J., Paquin, A.-C., Chan, M., Parisotto, K. L., Sarracino, J., & Gillis, R. L. (2011). Do we know when our impressions of others are valid? Evidence for realistic accuracy awareness in first impressions of personality. *Social Psychological and Personality Science*, *2*(5), 452–459. https://doi.org/10.1177/1948550610397211

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. https://doi.org/10.1037/0022-3514.86.4.599

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

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

Cope, M. (2010). Coding qualitative data. *Qualitative Research Methods in Human Geography*, *3*, 281–294.

Decristan, J., Kunter, M., & Fauth, B. (2022). Die Bedeutung individueller Merkmale und konstruktiver Unterstützung der Lehrkraft für die soziale Integration von Schülerinnen und Schülern im Mathematikunterricht der Sekundarstufe [The relevance of individual characteristics and teacher's constructive support for students' social integration in mathematics instruction in secondary education]. *Zeitschrift für Pädagogische Psychologie*, *36*(1–2), 85–100. https://doi.org/10.1024/1010-0652/a000329

Decristan, J., Kunter, M., Fauth, B., Büttner, G., Hardy, I., & Hertel, S. (2016). What role does instructional quality play for elementary school children's science competence? A focus on students at risk. *Journal for Educational Research Online*, *8*(1), 66–89. https://doi.org/10.25656/01:12032

de Freitas, E. (2015). The moving image in education research: Reassembling the body in classroom video data. *International Journal of Qualitative Studies in Education*, *29*(4), 553–572. https://doi.org/10.1080/09518398.2015.1077402

de Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on Psychological Science*, *16*(6), 1412–1427. https://doi.org/10.1177/1745691620964172

Desi-Konsortium (Eds.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* [Teaching and competency acquistion in German and English: Results of the DESI-study]. Beltz.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy*, *24*(2), 267–329. https://doi.org/10.1177/0895904808330173

Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, Practice, and Contemporary Issues* (pp. 97–125). Lawrence Erlbaum Associates Publishers.

Dresing, T., & Pehl, T. (2012). *Praxisbuch Transkription: Regelsysteme, Software und Anleitungen für qualitative ForscherInnen* (4th ed.) [Practical Guide to Transcription: Rule Systems, Software, and Instructions for Qualitative Researchers]. Eigenverlag.

Evans, J. S. B. T. (2006). Dual system theories of cognition: Some issues. *Proceedings of the Annual Meeting of the Cognitive Science Society*, (*28*), 202–207.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2018). Dual process theory: Perspectives and problems. In W. de Neys (Ed.), *Current Issues in Thinking and Reasoning. Dual Process Theory 2.0* (pp. 137–155). Routledge.

Evans, J. S. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fauth, B., Herbein, E., & Maier, J. L. (2024). *Beobachtungsmanual zum Unterrichtsfeedbackbogen Tiefenstrukturen* [Observation Manual for the Classroom Feedback Questionnaire on Deep-Structures]. Institut für Bildungsanalysen Baden-Württemberg.

Fowler, K. A., Lilienfeld, S. O., & Patrick, C. J. (2009). Detecting psychopathy from thin slices of behavior. *Psychological Assessment*, *21*(1), 68–78. https://doi.org/10.1037/a0014938

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

Gawronski, B., Luke, D. M., & Creighton, L. A. (2024). Dual-process theories. In D. E. Carlston, K. Hugenberg, & K. L. Johnson (Eds.), *The Oxford Handbook of Social Cognition* (2nd ed., pp. 319–353). Oxford University Press.

Göllner, R., Fauth, B., & Wagner, W. (2021). Student ratings of teaching quality dimensions: Empirical findings and future directions. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student Feedback on Teaching in Schools* (pp. 111–122). Springer International Publishing. https://doi.org/10.1007/978-3-030-75150-0_7

Goodell, L. S., Stage, V. C., & Cooke, N. K. (2016). Practical qualitative research strategies: Training interviewers and coders. *Journal of Nutrition Education and Behavior*, *48*(8), 578–585. https://doi.org/10.1016/j.jneb.2016.06.001

Hall, J. A., Horgan, T. G., & Murphy, N. A. (2019). Nonverbal Communication. *Annual Review of Psychology*, *70*(1), 271–294. https://doi.org/10.1146/annurev-psych-010418-103145

Hannafin, M. J., Shepherd, C. E., & Polly, D. (2010). Video assessment of classroom teaching practices: Lessons learned, problems and issues. *Educational Technology*, *59*(1), 32–37.

Hardy, I., Hertel, S., Kunter, M., Klieme, E., Warwas, J., Büttner, G., & Lühken, A. (2011). Adaptive Lerngelegenheiten in der Grundschule. Merkmale, methodisch-didaktische Schwerpunktsetzungen und erforderliche Lehrerkompetenzen [Adaptive learning opportunities in elementary school. Characteristics, methodological-didactic prioritization and required teacher competences]. *Zeitschrift für Pädagogik, 57(6)*, 819–833. https://doi.org/10.25656/01:8783

Hattie, J. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Helmke, A. (2014). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* [Teaching Quality and Teacher Professionalism: Diagnosis, Evaluation, and Improvement of Instruction]. Klett/Kallmeyer.

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. https://doi.org/10.1016/j.jrp.2009.03.017

Hyytinen, H., Holma, K., Toom, A., & Shavelson, R. J., & Lindblom-Ylänne, S. (2014). The complex relationship between students' critical thinking and epistemological beliefs in the context of problem solving. *Frontline Learning Research*, *2*(5), 1–25. https://doi.org/10.14786/flr.v2i4.124

Janik, T., & Seidel, T. (Eds.). (2013). *The Power of Video Studies in Investigating Teaching and Learning in the Classroom*. Waxman.

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

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2011). *Thinking fast and slow*. Macmillan.

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kawulich, B. B. (2004). Data analysis techniques in qualitative research. *Journal of Research in Education*, *14*(1), 96–113.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*(6), 533–550. https://doi.org/10.1111/j.1745-6924.2009.01164.x

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In J. Tomáš & T. Seidel (Eds.), *The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (pp. 137–160). Waxmann.

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology* (2nd ed.). Sage.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. https://doi.org/10.1037/a0020762

Kuger, S., Klieme, E., Lüdtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Mathematikunterricht und Schülerleistung in der Sekundarstufe: Zur Validität von Schülerbefragungen in Schulleistungsstudien [Mathematics instruction and student achievement in secondary education: About the validity of student's survey and school achievement studies]. *Zeitschrift für Erziehungswissenschaft*, *20*(2), 61–98. https://doi.org/10.1007/s11618-017-0750-6

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., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., Jordan, A., & Neubrand, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler: Schulformunterschiede in der Unterrichtsqualität [Mathematics instruction of the PISA-students]. *Zeitschrift für Erziehungswissenschaft*, *8*(4), 502–520. https://doi.org/10.1007/s11618-005-0156-8

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). Waxmann.

Lambert, N. M., Mulder, S., & Fincham, 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

Learnpulse SAS (2020). *Screenpresso PRO* (Version 1.8.3.0) [Computer software]. https://www.screenpresso.com/de/

Leighton, J. P. (2017). *Using Think-Aloud Interviews and Cognitive Labs in Educational Research*. Oxford University Press. https://doi.org/10.1093/9780199372904.001.0001

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

Marzano, R. J., & Marzano, J. S. (2003). The key to classroom management. *Educational Leadership*, *61*(1), 6–13.

McKim, C. A. (2017). The value of mixed methods research: A mixed methods study. *Journal of Mixed Methods Research*, *11*(2), 202–222. https://doi.org/10.1177/1558689815607096

Miles, M. B., & Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. Sage.

Milli, S., Lieder, F., & Griffiths, T. L. (2021). A rational reinterpretation of dual-process theories. *Cognition*, *217*, 104881. https://doi.org/10.1016/j.cognition.2021.104881

Mugg, J. (2015). Two minded creatures and dual-process theory. *Journal of Cognition and Neuroethics*, *3*(3), 87–112.

Murphy, N. A., & Hall, J. A. (2021). Capturing behavior in small doses: A review of comparative research in evaluating thin slices for behavioral measurement. *Frontiers in Psychology*, *12*, 667326. https://doi.org/10.3389/fpsyg.2021.667326

Namey, E., Guest, G., Thairu, L., & Johnson, L. (2008). Data reduction techniques for large qualitative data sets. In G. Guest, & K. MacQueen (Eds.), *Handbook for Team-Based Qualitative Research,* (pp. 137–161). Rowman & Littlefield.

Nesse, R. M. (2005). Natural selection and the regulation of defenses. *Evolution and Human Behavior*, *26*(1), 88–105. https://doi.org/10.1016/j.evolhumbehav.2004.08.002

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*(2), 291–310. https://doi.org/10.1037/0033-295X.108.2.291

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1–13. https://doi.org/10.1177/1609406919899220

Pennycook, G. (2017). A Perspective on the theoretical foundation of dual process models. In W. De Neys (Ed.), *Dual Process Theory 2.0* (1st ed., pp. 5–27). Routledge. https://doi.org/10.4324/9781315204550-2

Petko, D., Waldis, M., Pauli, C., & Reusser, K. (2003). Methodologische Überlegungen zur videogestützten Forschung in der Mathematikdidaktik: Ansätze der TIMSS 1999 Video Studie und ihrer schweizerischen Erweiterung [Methodological considerations about video-based research in mathematics didactic]. *Zentralblatt für Didaktik der Mathematik*, *35*(6), 265–280. https://doi.org/10.1007/BF02656691

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

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

Pretsch, J., Flunger, B., Heckmann, N., & Schmitt, M. (2013). Done in 60 s? Inferring teachers' subjective well-being from thin slices of nonverbal behavior. *Social Psychology of Education*, *16*(3), 421–434. https://doi.org/10.1007/s11218-013-9223-9

Rädiker, S., & Kuckartz, U. (2019). *Analyse qualitativer Daten mit MAXQDA: Text, Audio und Video* [Analysis of qualitative data with MAXQDA: Text, Audio, and Video]. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-22095-2

Rimondini, M., Mazzi, M. A., Busch, I. M., & Bensing, J. (2019). You only have one chance for a first impression! Impact of patients' first impression on the global quality assessment of doctors' communication approach. *Health Communication*, *34*(12), 1413–1422. https://doi.org/10.1080/10410236.2018.1495159

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, *29*(8), 1358–1369. https://doi.org/10.1177/0956797618774253

Saldaña, J. (2009). *The Coding Manual for Qualitative Researchers*. Sage.

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research*, *3*(3), 208–222. https://doi.org/10.1177/1558689809334210

Schensual, J. J., LeCompte, M. D., Hess; G. A., Nastasi, B. K., Berg, M. J., Williamson, L., Brecher, J. & Glasser, R. (1999). *Using Ethnographic Data: Interventions, Public Programming and Public Policy* (Vol. 7). AltaMira.

Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design. *KZfSS Kölner Zeitschrift Für Soziologie und Sozialpsychologie*, *69*(S2), 107–131. https://doi.org/10.1007/s11577-017-0454-1

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Sokolovic, N., Brunsek, A., Rodrigues, M., Borairi, S., Jenkins, J. M., & Perlman, M. (2021). Assessing quality quickly: Validation of the Responsive Interactions for Learning – Educator (RIFL-Ed.) measure. *Early Education and Development*, *33*(6), 1061–1076. https://doi.org/10.1080/10409289.2021.1922851

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *Two Minds: Dual Processes and*

Beyond (1st ed., pp. 55–88). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stanovich, K. E., West, R. F., & Toplak, M. E. (2014). Rationality, intelligence, and the defining features of Type 1 and Type 2 processing. In J. W. Sherman, B. Gawronski & Y. Trope (Eds.), *Dual-Process Theories of the Social Mind* (pp. 80–91). The Guilford Press.

Tackett, J. L., Herzhoff, K., Kushner, S. C., & Rule, N. (2015). Thin slices of child personality: Perceptual, situational, and behavioral contributions. *Journal of Personality and Social Psychology*, *110*(1), 150–66. https://doi.org/10.1037/pspp0000044

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The new era of mixed methods. *Journal of Mixed Methods Research*, *1*(1), 3–7. https://doi.org/10.1177/2345678906293042

Trautwein, U., Sliwka, A., & Dehmel, A. (2022). *Grundlagen für einen wirksamen Unterricht. Reihe wirksamer Unterricht Band 1* [Basics of Effective Teaching. Series on Effective Teaching, Vol. 1]. Institut für Bildungsanalysen Baden-Württemberg.

Van Den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351. https://doi.org/10.1080/0044929031000

Vinokic, K., Baron, F., Kunter, M., Linberg, A., Begrich, L., & Kuger, S. (2024). Using the thin slices technique to assess interactional quality in early childhood education and care settings. *Frontiers in Education*, *9*, 1368503. https://doi.org/10.3389/feduc.2024.1368503

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

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721. https://doi.org/10.1037/edu0000075

White, M., & Ronfeldt, M. (2024). Monitoring rater quality in observational systems: Issues due to unreliable estimates of rater quality. *Educational Assessment*, *29*(2), 124–146. https://doi.org/10.1080/10627197.2024.2354311

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592-598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education*, *19*(3), 409–427. https://doi.org/10.1007/s10459-013-9453-9

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(1), 51–68. https://doi.org/10.1002/acp.1331

Appendix

The questions of the guided interview

1. At first, please tell us with regard to all videos and all questionnaires: What was going on in your head while you were watching the videos and completing the questionnaires?
2. How did you arrive at your judgment? How did you proceed?
3. What was particularly difficult or easy?
4. How did your decision evolve? Was it rather reflective or was it rather a gut decision?
5. Did some impressions influence your ratings in particular?
6. What was crucial for your decision? Was it rather the video or rather the questionnaire or both?
7. Now let us talk about the teachers: Was something noticeable about the teachers, which was specifically remarkable?
8. What about the audio content? How important was it?
9. On which sector of the screen did you focus?
10. Did you pay attention to the gesture, posture, voice or clothes of the teacher? If so, how did it influence your ratings?
11. Did you notice a smile? If so, how did it influence your decision?
12. In how far was the attitude of the students, the atmosphere or the condition of the classroom crucial for your ratings?