OPEN ACCESS
OPUS AUGSBURG

Universität Augsburg
Universitätsbibliothek

Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction

Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, Ulrich Trautwein

Angaben zur Veröffentlichung / Publication details:

Goldberg, Patricia, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2021. "Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction." *Educational Psychology Review* 33 (1): 27–49. https://doi.org/10.1007/s10648-019-09514-z.

Nutzungsbedingungen / Terms of use: **CC BY 4.0**

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Educational Psychology Review (2021) 33:27–49
https://doi.org/10.1007/s10648-019-09514-z

REVIEW ARTICLE

Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction

Patricia Goldberg[1] · Ömer Sümer[1] · Kathleen Stürmer[1] · Wolfgang Wagner[1] · Richard Göllner[1] · Peter Gerjets[2] · Enkelejda Kasneci[3] · Ulrich Trautwein[1]

Published online: 18 December 2019
© The Author(s) 2019

Abstract
Teachers must be able to monitor students' behavior and identify valid cues in order to draw conclusions about students' actual engagement in learning activities. Teacher training can support (inexperienced) teachers in developing these skills by using videotaped teaching to highlight which indicators should be considered. However, this supposes that (a) valid indicators of students' engagement in learning are known and (b) work with videos is designed as effectively as possible to reduce the effort involved in manual coding procedures and in examining videos. One avenue for addressing these issues is to utilize the technological advances made in recent years in fields such as machine learning to improve the analysis of classroom videos. Assessing students' attention-related processes through visible indicators of (dis)engagement in learning might become more effective if automated analyses can be employed. Thus, in the present study, we validated a new manual rating approach and provided a proof of concept for a machine vision-based approach evaluated on pilot classroom recordings of three lessons with university students. The manual rating system was significantly correlated with self-reported cognitive engagement, involvement, and situational interest and predicted performance on a subsequent knowledge test. The machine vision-based approach, which was based on gaze, head pose, and facial expressions, provided good estimations of the manual ratings. Adding a synchrony feature to the automated analysis improved correlations with the manual ratings as well as the prediction of posttest variables. The discussion focuses on challenges and important next steps in bringing the automated analysis of engagement to the classroom.

Keywords Students' visible engagement · Attention-related behavior · Machine learning · Automated picture analysis · Classroom synchronization

✉ Patricia Goldberg
patricia.goldberg@uni-tuebingen.de

Extended author information available on the last page of the article

Cognitive activation, classroom management, and teacher support are the three central tenants of teaching quality (Klieme et al. 2006; Praetorius et al. 2018). The level of students' (dis)engagement in learning activities can be considered a major indicator of both cognitive activation and classroom management because it signals students' engagement in the deep processing of learning content and reveals the time on task (Carroll 1963) provided by the teachers for students' learning. To this end, teachers are required to take note of their students' attentional focus and make sure the students are engaging in the desired learning activities. Thus, the ability to monitor students' attention and to keep it at a high level is part of the competencies that novice teachers need to acquire. However, research has indicated that teachers might not always be aware of their students' attentional focus, and this may be particularly true for novice teachers.

In general, beginning teachers have trouble monitoring all students in the classroom evenly and noticing events that are relevant for student learning (Berliner 2001; Cortina et al. 2015; Star and Strickland 2008; Stürmer et al. 2017). Therefore, teacher training needs to support future teachers in developing the necessary knowledge structures that underlie these abilities (e.g., Lachner et al. 2016). Consequently, providing an improved measurement approach for student attention will be beneficial for research and can potentially contribute to teacher training. Research has already demonstrated that both inexperienced and experienced teachers' ability to notice relevant cues in the classroom benefits from observing and reflecting on their own videotaped teaching (Kleinknecht and Gröschner 2016; Sherin and van Es 2009). Until now, however, instructors have typically had to watch hours of video material to select the most crucial phases of lessons. Similarly, when it comes to research on teaching effectiveness and the development of teachers' ability to notice relevant cues in classroom instruction (i.e., professional vision skills), researchers typically have to invest considerable resources, especially coding resources, to examine the association between teacher behavior and classroom processes (Erickson 2007). The required effort further increases when investigating students' attention across an entire lesson and analyzing attention at the group level instead of among individuals. In this vein, attention- and engagement-related behavior during classroom instruction has rarely been studied due to the difficulty of data collection and labeling. However, learners might behave differently in naturalistic settings and show versatile behavior that cannot be found in a lab.

One potentially valuable avenue for addressing these issues is to utilize the technological advances made in recent years in fields such as computer vision and machine learning. Therefore, in an ongoing research project (Trautwein et al. 2017), we have been investigating whether and how the automated assessment of students' attention levels can be used as an indicator of their active engagement in learning. This automated assessment can in turn be used to report relevant cues back to the teacher, either simultaneously or by identifying and discussing the most relevant classroom situations (e.g., a situation where students' attention increases or decreases significantly) after a lesson.

In the present study, we present a proof of concept for such a machine vision-based approach by using manual ratings of visible indicators of students' (dis)engagement in learning as a basis for the automated analysis of pilot classroom recordings of three lessons with university students. More specifically, by combining multiple indicators from previous research (i.e., Chi and Wylie 2014; Helmke and Renkl 1992; Hommel 2012), we developed a manual rating instrument to continuously measure students' observable behavior. In addition, we performed an automated analysis of the video recordings to extract features of the students' head pose, gaze direction, and facial expressions using modern computer vision techniques. Using these automatically extracted features, we aimed to estimate manually annotated

attention levels for each student. Because we had continuous labeling, this could be done by training a regressor between the visible features and the manual labels. We investigated the predictive power of both the manual and automatic analyses for learning (i.e., performance on a subsequent knowledge test). To account for complexity within classrooms and enrich the automated analysis, we also considered synchronous behavior among neighboring students. In the present article, we report initial empirical evidence on the reliability and validity of our automated assessments and their association with student performance.

Attention in Classroom Instruction

Student attention is a key construct in research on both teaching and learning. However, definitions vary widely and are discussed from multiple perspectives. Here, we focus on describing three lines of research that inspired our research program: cognitive psychology models that describe attention as part of information processing, engagement models in which attention makes up part of a behavioral component, and teaching quality models in which student attention is a crucial factor.

In current models in the psychology of learning, *attention* denotes a filtering mechanism that determines the kind and amount of information that enters working memory (Driver 2001). This mechanism is crucial for preventing working memory overload and allows the learner to focus on the right kind of information. Only sensory information that enters working memory is encoded, organized, and linked to already existing knowledge. Thus, attention serves as a selection process for all incoming sensory information as it dictates which pieces of information will be processed further and will get the chance to be learned. Thus, attention determines the success of knowledge construction (Brünken and Seufert 2006). Engle (2002) further proposed that executive attention, which actively maintains or suppresses current representations in working memory, is part of working memory. Certain instructional situations strongly depend on executive processes such as shifting, inhibition, or updating (Miyake et al. 2000) and thus necessitate top-down attentional control. Although information processing occurs in a covert manner, some aspects of attentional processes are likely to be observed from the outside: for example, visually orienting toward a certain stimulus, which improves processing efficiency (Posner 1988).

Attention is often mistaken for engagement, even though it constitutes only part of it. *Engagement* is defined as a multidimensional meta-construct and represents one of the key elements for learning and academic success (Fredricks et al. 2004). It includes observable behaviors, internal cognitions, and emotions. Covert processes such as investment in learning, the effort expended to comprehend complex information, and information processing form part of cognitive engagement (Fredricks et al. 2004; Pintrich and De Groot 1990). Emotional engagement in the classroom includes affective reactions such as excitement, boredom, curiosity, and anger (Connell 1990; Fredricks et al. 2004). Attention is considered a component of behavioral engagement alongside overt participation, positive conduct, and persistence (Connell 1990; Fredricks et al. 2004). Per definition, cognitive engagement refers to internal processes, whereas only the emotional and behavioral components are manifested in visible cues. Nevertheless, all engagement elements are highly interrelated and do not occur in isolation (Fredricks et al. 2004). Thus, attention plays a crucial role because it may signal certain learning-related processes that should become salient in students' behavior to some extent.

Learners' attention also plays a crucial role in *research on teaching*. Teachers must determine whether their students are attentive by considering visible cues, continually monitoring the course of events in order to manage the classroom successfully (Wolff et al. 2016) and providing ambitious learning opportunities. A student's attention or lack thereof (e.g., when distracted or engaging in mind wandering) can signal whether she or he is on-task or off-task. This in turn can provide hints about instructional quality and the teacher's ability to engage his or her students in the required learning activities. Thus, it is important to help teachers develop the skills needed to monitor and support student attention and engagement and adapt their teaching methods. Consequently, accounting for student attention and more broadly student engagement in teaching is considered crucial for ensuring teaching quality, including classroom management, cognitive activation, and instructional support (Klieme et al. 2001; Pianta and Hamre 2009).

In sum, the definitions, theoretical backgrounds, and terminology used in various lines of research to describe observable aspects of students' cognitive, affective, or behavioral attention/engagement in learning are diverse, but experts agree on their importance and key role in learning. As teachers must rely on visible cues to judge their students' current attention levels (Büttner and Schmidt-Atzert 2004; Yamamoto and Imai-Matsumura 2012), we focused on observable aspects of attention and inferences that were based on visible indicators. In the remainder of the article, we use the term *visible indicators of (dis)engagement in learning* to describe these aspects. These visible indicators are highly likely to be associated with learning, but this assumption needs to be validated.

Previous Approaches for Measuring Visible Indicators of Engagement in Learning

The difficulty in assessing students' engagement-related processes in real-world classroom settings consists of externalizing learners' internal (covert) states through visible overt aspects to the greatest extent possible. In psychology, affective states and cognitive processes such as attentional control are usually determined from physiological signals, such as heart rate, electrodermal activity, eye tracking, or electroencephalography (e.g., Gerjets et al. 2014; Krumpe et al. 2018; Poh et al. 2010; Yoshida et al. 2014). Using this kind of psychologically sound measurements makes it possible to detect covert aspects of learning-related processes; however, these measures are hardly feasible in classroom instruction, especially when teachers must be equipped with knowledge about what indicators to look for in students. Furthermore, these approaches are useful for answering very specific research questions. However, they are not sufficient for determining whether students' ongoing processes are actually the most appropriate for the situation. By contrast, overt behavior can provide visible indicators of appropriate learning-related processes in students.

Overt classroom behavior is an important determinant of academic achievement (Lahaderne 1968; McKinney et al. 1975). Although overt behavior does not always represent a reliable indicator of covert mental processes, previous findings have demonstrated a link between cognitive activity and behavioral activity (Mayer 2004). Previous studies have analyzed students' behavior and have determined its relation to achievement (Helmke and Renkl 1992; Hommel 2012; Karweit and Slavin 1981; Stipek 2002). Furthermore, in research on engagement, correlations between student engagement and academic achievement have been found (Lei et al. 2018). Other studies have found opposing results (e.g., Pauli and Lipowsky 2007); however, these

studies either relied on self-reports as opposed to observer ratings or only focused on certain facets of engagement-related behavior (e.g., only active on-task behavior).

There have been various attempts to systematically assess visible indicators of engagement in classroom learning, for example, Helmke and Renkl (1992) based their research on an idea by Ehrhardt et al. (1981) and related observable student behavior to internal processes using time-on-task as an indicator of whether a student was paying attention to classroom-related content. Assessing observable content-related behavior is essential to this operationalization of higher order attention. Hommel (2012) modified this approach and applied it to the video-based analysis of instructional situations. Rating behavior as either on- or off-task with varying subcategories demonstrated the interrelation between visual cues and achievement or reduced learning (Baker et al. 2004; Helmke and Renkl 1992).

However, learners can differ in their learning activities but still be engaged in a certain task. The ICAP framework proposed by Chi and Wylie (2014) distinguishes between passive, active, constructive, and interactive overt behavior, which differ across various cognitive engagement activities. This framework focuses on the amount of cognitive engagement, which can be detected from the way students engage with learning materials and tasks (Chi and Wylie 2014). This theoretical model provides a promising approach for further expanding the different types of on-task behavior so that variations in student behavior can be accounted for.

In sum, considering learning content has been shown to be useful; however, there is a lack of research involving the continuous analysis of attention or engagement over the course of one or more lessons. A unique feature of the present study is that we aimed to acquire a continuous assessment (i.e., a score for every student in the classroom for every second of instruction time) of students' visible indicators of (dis)engagement in learning. This temporal resolution was crucial in our approach because we aimed to provide comparable data that could be used to train a machine-learning algorithm. To reach this high level of temporal resolution, we decided to annotate learners' behavior continuously. The free software CARMA (Girard 2014) enables the continuous interpersonal behavior annotation by using joysticks (see Lizdek et al. 2012). However, this new approach limited us in terms of using already existing rating instruments because existing instruments do not allow for a high enough level of temporal resolution. Furthermore, the CARMA software requires annotations on a scale rather than rating the behavior in terms of categories as already existing instruments do. When developing the new instrument, we mainly oriented on the MAI (Helmke and Renkl 1992; Hommel 2012). However, we needed to define more fine-grained indicators of student behavior to make annotations along a continuous scale possible. Therefore, we added indicators from various established instruments to extend our rating scale. We assumed that the manual observer annotations would serve only as approximations of the actual cognitive states of the students and that the averaged (i.e., intersubjective) manual annotations would reflect the "true score" of the visible indicators of (dis)engagement in learning better than a single rater could. Subsequent to the ratings, we thus calculated the mean of the raters for every second. The mean values for each second and student were used as the ground truth to train a machine-learning approach.

Using Machine Learning to Assess Visible Indicators of (Dis)Engagement in Learning

Machine learning and computer vision methods have made tremendous progress over the past decade and have been successfully employed in various applications. In the context of

teaching, these methods might offer an efficient way to measure student engagement, thereby decreasing the need for human rating efforts. However, any machine-learning method that is aimed at estimating covert engagement-related processes in learning needs to depend on visible indicators such as head pose, gaze direction, facial action unit intensity, or body pose and gestures. State-of-the-art methodologies for the automated assessment of engagement can be divided into two categories: single-person- and classroom-based analyses.

In a single-person analysis, facial expressions can provide hints about ongoing cognitive processes and can be analyzed by considering action unit (AU) features. Related studies by Grafsgaard et al. (2013) and Bosch et al. (2016a, b) investigated the relations between AU features and several response items and affective states. Even though these studies found that several facial AUs were associated with engagement, they were limited to affective features and did not consider head pose or gaze direction.

In another work, Whitehill et al. (2014) introduced a facial analysis approach to estimating the level of engagement on the basis of manually rated engagement levels. Although their facial analysis approach was able to predict learning just as accurately as participants' pretest scores could, the correlation between engagement and learning was moderate due to the limited amount of data and the short-term nature of the situations.

In a classroom-based analysis, the focus shifts away from single individuals onto shared features and interactions among participants. In this context, a number of notable contributions (e.g., Raca 2015; Raca and Dillenbourg 2013) have utilized various sources of information to understand features of audience behavior, such as the amount of estimated movement and synchronized motions among neighboring students. They found that immediate neighbors had a significant influence on a student's attention, whereas students' motion was not directly connected with reported attention levels (Raca and Dillenbourg 2013; Raca et al. 2013). Furthermore, Raca et al. (2014) analyzed students' reaction time upon presentation of relevant information (sleeper's lag). In addition to estimating head pose, they considered the class period, student's row, how often faces were automatically detected (as a precursor to eye contact), head movement, and the amount of still time (i.e., 5-s periods without head movement) because these features had previously been shown to be good predictors of engagement in learning (Raca et al. 2015). Although these results were promising, they were limited to correlational studies of reported attention levels; predictive approaches were not used due to limits in the performance of computer vision methodology.

A recent study estimated human-annotated attention levels by using 3D vision cameras to identify individuals using face and motion recognition without any physical connection to people and solely on the basis of visual features (Zaletelj 2017; Zaletelj and Košir 2017). Due to technological limitations associated with 3D vision cameras, the analysis was based on a single row of students rather than the entire classroom. Fujii et al. (2018) used head-up and head-down states and classroom synchronization in terms of head pose as informative tools that could provide feedback to teachers. However, they did not validate their system using educational measures (pretests, posttests, or observations) and only reported user experiences with three teachers.

In sum, few previous studies have investigated classroom-based attention and engagement beyond the single-person context due to the poor performance of computer vision approaches for face and body pose recognition in unconstrained settings (e.g., varying illumination, occlusion, motion, challenging poses, low resolution, and long distance). However, recent advances in deep learning technology have resulted in the availability of new methods for the robust extraction of such features from videos. By employing such technology in this study, we

aim to bring a fine-scaled analysis of visible indicators to classroom studies and augment individual engagement analysis with another useful feature: classroom synchronization.

Research Questions

The present study is part of an ongoing research project in which researchers from education science, psychology, and computer science are working to create an automatic assessment of students' engagement that could one day be implemented in an interface that can be used for research as well as teacher training purposes. The present study lays the basis for achieving these goals by developing and testing an automated approach to assessing visible indicators of students' (dis)engagement in learning. Such a remote approach requires comparable data (generated by human raters) that can be used as the ground truth in order to train a classifier. However, existing instruments (Helmke and Renkl 1992; Hommel 2012) for measuring engagement-related processes in learning (a) require human observers to make a huge number of inferences and (b) require data to be collected in 30-s or 5-min intervals. This is problematic for our context because an automated analysis can only rely on visible indicators, does not consider content-specific information at all, and operates at a more fine-grained temporal resolution. Therefore, we developed a new instrument to annotate student behavior manually by applying a rating method with visible indicators over time. This manual rating served as the starting point from which to train an algorithm by applying methods from machine learning and computer vision.

The present study addressed the following research questions:

1) Is the new manual annotation of visible indicators of (dis)engagement in learning related to students' learning processes and outcomes? To validate our instrument, we examined how the manual ratings were correlated with students' self-reported cognitive engagement, involvement, and situational interest. We expected these self-reported learning activities to cover different facets of (dis)engagement in learning, and when combined, we expected them to account for cognitive parts of the construct. Furthermore, we tested whether the scores resulting from the manual annotation would predict students' performance on a knowledge test at the end of an instructional session.
2) Is it possible to adequately replicate the relation to students' learning processes and outcomes by using visible indicators of (dis)engagement in learning based on the machine-learning techniques that estimated the manual ratings? We used gaze, head posture, and facial expressions to estimate the manual ratings. To test the quality of our machine vision-based approach, we examined the associations between the scores generated from the automated approach and the manual ratings and students' self-report data regarding their learning processes, and we used the machine-learning scores to predict achievement on the knowledge test.
3) How do adding synchrony aspects of student behavior affect the automated estimations of the manual ratings? The results of previous studies have indicated that immediate neighbors have a significant influence on a student's engagement (Raca and Dillenbourg 2013; Raca et al. 2013). As a first step toward including indicators of synchrony in our project, we added students' synchrony with the person sitting next to them as an additional variable to our prediction models, which were based on the automated assessment of student engagement.

Method

The ethics committee from the Leibniz-Institut für Wissensmedien in Tübingen approved our study procedures (approval #2018-017), and all participants gave written consent to be videotaped.

Sample and Procedure

We decided to conduct a study involving university students in order to validate our approach before administering it in school classrooms. A total of $N = 52$ university students (89.5% women, 8.8% men, mean age = 22.33, $SD = 3.66$) at a German university volunteered to take part in the study. The study was conducted during regular university seminar sessions on quantitative data analysis (90 min). A total of three different seminar groups were assessed. The topics of the sessions were either *t tests for independent samples* (sessions 1 and 2) or *regressions* (session 3) and ranged from 30 to 45 min. The sessions were videotaped with three cameras (one teacher camera, two cameras filming the students). If students refused to be videotaped, they were either seated outside the scope of the cameras or switched to a parallel seminar. Participants were informed in advance of the study's purpose, procedure, and ethical considerations such as data protection and anonymization. To avoid confounding effects of the teacher, the same person taught all sessions in a teacher-centered manner. Before the session started, students filled out a questionnaire on background variables (age, gender, final high school examination [Abitur] grade, school type) and individual learning prerequisites. After the session, participants completed a knowledge test on the specific topic of the session and completed another questionnaire about learning activities during the seminar.

Instruments

Individual Learning Prerequisites We used established questionnaire measures to assess three individual learning prerequisites: Dispositional interest in the session's topic was captured with four items ($\alpha = .93$) adapted from Gaspard et al. (2017). Self-concept in quantitative data analysis was assessed with five items ($\alpha = .80$; adapted from Marsh et al. 2006), and 13 items were used to test for self-control capacity ($\alpha = .83$; Bertrams and Dickhäuser 2009). Moreover, we administered the short version of the quantitative subscale (Q3) of the cognitive abilities test (Heller and Perleth 2000). Measuring these learning prerequisites allowed us to control for potential confounding variables in the analyses.

Learning Outcomes The knowledge test consisted of 12 and 11 items that referred to participants' declarative and conceptual knowledge of the session topic, respectively. We *z*-standardized the knowledge test scores within each group for subsequent analysis.

Self-Reported Learning Activities After the session, we assessed students' involvement (four items, $\alpha = .61$; Frank 2014), cognitive engagement (six items, $\alpha = .79$; Rimm-Kaufman et al. 2015), and situational interest (six items, $\alpha = .89$; Knogler et al. 2015) during the seminar session (see Table 1).

Analysis

Continuous Manual Annotation To develop a continuous manual annotation that included potential valid indicators of students' visible (dis)engagement in learning, we used the instruments developed by Helmke and Renkl (1992) and Hommel (2012) as a basis. However, these instruments label behavior in categories and thus cannot be used as a continuous scale. Therefore, we combined the idea of on-/off-task behavior and active/passive subcategories with existing scales from the engagement literature. Furthermore, we used the theoretical assumptions about students' learning processes and related activities in classrooms pointed out by the ICAP framework (Chi and Wylie 2014) as an inspiration to define more fine-grained differentiations within the possible behavioral spectrum. The distinction into passive, active, constructive, and interactive behavior allowed us to make subtler distinctions between the different modes of on-task behavior, and this concept could be transferred to off-task behavior (i.e., passive, active, deconstructive, and interactive) as well. By combining different approaches, we could define visible indicators of (dis)engagement in learning on a continuous scale. The resulting scale ranged from −2, indicating interruptive and disturbing off-task behavior, to +2, indicating highly engaged on-task behavior where, for example, learners ask questions and try to explain the content to fellow learners (see Fig. 1). When a person could not be seen or was not present in the classroom, the respective time points were coded as missing values in subsequent analyses.

The behavior of each observed person throughout the instructional session was coded in 1-s steps using the CARMA software (Girard 2014) and a joystick. A total of six raters annotated the videotaped seminar sessions, and each session was annotated by a total of three raters. The raters consisted of student assistants and one researcher, all of whom were trained carefully before annotating the videos. First, raters were introduced to the conceptual idea of the rating and the rating manual.

Table 1 Item wording for learning activities

Construct	Items
Cognitive engagement	I exerted myself as much as possible during the session.
	I thought about different things during the session.
	I only paid attention when it was interesting during the session.
	It was important for me to really understand things during the session.
	I tried to learn as much as possible during the session.
	I pondered a lot during the session.
Involvement	During the session…
	… I strongly concentrated on the situation.
	… I occasionally forgot that I was taking part in a study.
	… I was mentally immersed in the situation.
	… I was fully engaged with the content.
Situational interest	When you think about today's session…
	… the seminar session aroused your curiosity.
	… the seminar session attracted your attention.
	… you were completely concentrated on the seminar session.
	… the seminar session was entertaining for you.
	… the seminar session was fun for you.
	… the seminar session was exciting for you.

Items have been translated

They were told to concentrate on observable behavior to avoid making inferences and considering information from previous ratings. The raters focused on one student at a time in a random order. Every rater had to code one of two specific sections of the video for training, and the raters had to annotate special students who showed different types of behavior. To ensure that we could use all the video material for our analysis, raters who used video section A for training annotated video section B later and vice versa. The respective video sections used for training purposes were not included in the analysis. Only after their annotations reached an interrater reliability with an expert rating of at least ICC(2,1) = .60 were raters allowed to annotate the study material. We report the ICC(2,1) here as an indicator of interrater reliability because our data were coded on a metric scale level, and we had more than two raters per participant. We calculated the ICC(2,1) for every student, indicating the interrater reliability averaged across all time points, whereby values between .60 and .74 indicated good interrater reliability (Hallgren 2012); the ICC(2,1) for each student was .65 on average (absolute agreement). When the annotations between the raters deviated strongly, critical situations were discussed among the raters and recoded following consensus. The raters were not informed about the students' individual prerequisites, their learning outcomes, or their self-reported learning activities.

Machine-Learning Approach In addition to the manual ratings (see previous section), we employed a machine vision-based approach to estimate (dis)engagement in learning using visible indicators and analyzed the same videos with this approach. More specifically, we first detected the faces in the video (Zhang et al. 2017) and automatically connected the faces detected in the video stream to each student so that we could track their behavior. Faces were aligned, and their representative features extracted automatically based on the OpenFace library (Baltrušaitis et al. 2018). However, this

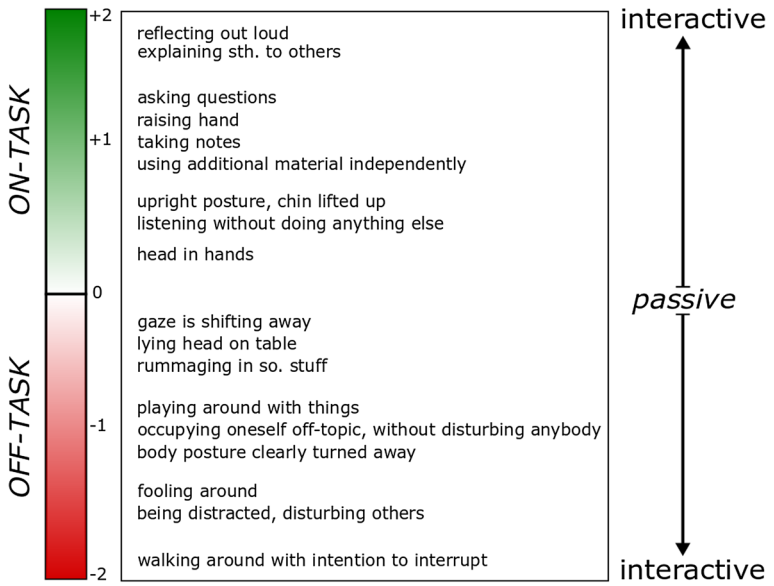

Fig. 1 Scale with exemplary behavioral indicators

procedure was not applicable to all students and all frames due to occlusions by peers, laptops, or water bottles. The subsequent analyses were therefore based on a subsample of $N = 30$ students.

In contrast to typical facial analysis tasks such as face recognition, the number of participants in classrooms is limited. We used the following three modalities as feature representations: *head pose*, *gaze direction*, and *facial expressions* (represented by facial action units). The head pose features consist of the head's location with respect to the camera and the rotation in radians around three axes. Gaze is represented by unit gaze vectors for both eyes and gaze direction in radians in world coordinates. Facial action units (AU) were estimated according to the Facial Action Coding System (FACS; Ekman and Friesen 1978), for which each AU can be expressed at five intensity levels. More specifically, to estimate the occurrence and intensity of FACS AUs, we used the following 17 AUs: upper face AUs are AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), and AU7 (lid tightener); the lower face AUs are AU9 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU23 (lip tightener), AU25 (lips part), AU26 (jaw drops), and AU45 (blink). Given that our videos were recorded at 24 frames per second, and the manual annotations were conducted each second, we used the mean values of these features for time sequences of 24 frames to predict engagement intensities. More specifically, we regressed the engagement intensities using linear Support Vector Regression (Fan et al. 2008) in a subject-independent manner. Excluding the subject whose engagement intensity was to be predicted, individual regression models were trained using all other student features and labels. Subsequently, the test subject's engagement during each 1-s period was predicted. Finally, the average estimated engagement intensity during the instructional session was taken as the final descriptor for each participant.

The label space for students' manually annotated engagement was between -2 and $+2$; however, the distribution of the data was highly imbalanced. Nearly 80% of all of the annotated data ranged from 0.2 to 0.8. Therefore, we had to clip the label values to fit the range of -0.5 and 1.5 and then rescale them to 0 and 1 in our regression models.

In summary, the visible indicators we used could be differentiated into two categories: engagement-related features (i.e., head pose and gaze direction) and emotion-related features (AU intensities). In order to compare their contributions with visible indicators of (dis)engagement in learning, we used them both separately and in combination.

In order to go beyond a single-person analysis, we further integrated an indicator of *synchrony*. Because simultaneous (i.e., synchronous) behavior in a group of students or an entire classroom can have an impact on individual students, in this first step toward an automated approach, we considered the behavior of neighboring students sharing the same desk. First, we measured the cosine similarities between neighboring students' manual ratings ($N = 52$, 26 pairs). Second, we calculated the relation between neighbors' synchrony (cosine similarities) and their mean engagement levels during instruction. Because synchronization is a precursor to engagement, we expected the neighbors to provide valuable information for estimating (dis)engagement in learning. Therefore, in the final step of our analysis, we concatenated the feature vector of each student and his or her neighbor into a single vector and trained the same regression models as for the estimation of each individual student's engagement.

Results

Relation Between Continuous Manual Annotation and Student Learning

We tested the validity of our manual rating instrument in two steps. First, we investigated construct validity by correlating the manual ratings with the self-reported learning activities. The manual annotations were significantly correlated with students' self-reported cognitive engagement, situational interest, and involvement ($.49 \leq r < .62$; Table 2).

Additionally, we calculated a multiple linear regression with the three self-reported learning activities as regressors. Together, they explained 42.9% of the variance in the manual ratings. This corresponds to a multiple correlation of $r = .66$. Second, we examined the predictive validity of our new instrument. We inspected the intercorrelations between all variables with the knowledge test (Table 2). The knowledge test scores (the dependent variable in this study) were significantly correlated with the manual ratings, cognitive abilities, and situational interest ($.30 \leq r < .42$). To test for effects of possible confounding variables, we calculated two additional linear regression models in which we added background variables (model 2) and learning prerequisites (model 3) into the regression and compared them with the prediction that involved only manual ratings (Table 3). The effect of the manual ratings remained robust and still explained a significant proportion of the variance in the knowledge test results.

Reanalysis with Machine-Learning Approach

We applied our trained regression to test subjects at 1-s intervals and applied mean pooling to create a final estimation that summarized participants' engagement. Table 4 shows the performance of different modalities for estimating (dis)engagement in learning. The performance measures were mean squared errors in the regression and the Pearson correlation coefficient between the manual annotations' mean level and our models' prediction during the instructional session.

As shown in Table 4, the head pose modality exhibited a lower correlation with the manual ratings ($r = .29$) than the other features. By contrast, gaze information and facial expressions (AU intensities) were more strongly correlated with the manual annotations ($r = .44$). Combining head pose and gaze ($r = .61$) or all three modalities ($r = .61$) also led to substantial correlations with the manual annotations.

In addition, we tested the correlations between the posttest variables (i.e., the knowledge test and self-reported learning activities) and the different models for estimating the manual ratings (Table 5). According to these results, regression models, which perform better with respect to MSE and lead to higher correlations with the manual ratings, seem to contain more information that is relevant for the posttest variables, particularly with respect to involvement and cognitive engagement.

Addition of Synchrony to the Machine-Learning Approach

The cosine similarities of the manual annotations between neighboring students were strongly correlated with each neighbor's mean engagement level throughout the recording ($r = .78$). More specifically, taking the synchronization into consideration improved the correlation with the manual ratings by 9%, thus showing that synchronization information is helpful for understanding (dis)engagement in learning.

Table 2 Correlations between individual characteristics, learning activities, achievement, and manual rating, with confidence intervals in brackets

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Female												
2. Age	−.29* [−.52, −.01]											
3. Abitur grade	−.10 [−.36, .19]	.31* [.03, .54]										
4. School type	−.09 [−.36, .20]	−.01 [−.29, .27]	−.26 [−.51, .01]									
5. Dispositional interest	−.12 [−.39, .16]	.04 [−.24, .32]	−.07 [−.34, .21]	.01 [−.27, .29]								
6. Self-concept	.16 [−.12, .42]	−.22 [−.47, .06]	.16 [−.12, .42]	.00 [−.28, .28]	−.62** [−.77, −.41]							
7. Self-control capacity	.09 [−.20, .36]	−.10 [−.37, .18]	−.15 [−.41, .14]	−.14 [−.41, .14]	.28 [−.00, .52]	−.38** [−.60, −.12]						
8. Cognitive abilities	.04 [−.24, .31]	.07 [−.22, .34]	−.39** [−.60, −.12]	.08 [−.21, .35]	.14 [−.15, .40]	−.22 [−.47, .06]	−.02 [−.30, .26]					
9. Manual rating	−.21 [−.46, .08]	.04 [−.24, .32]	.02 [−.26, .30]	−.25 [−.49, .03]	.18 [−.10, .44]	−.21 [−.46, .07]	.20 [−.08, .45]	.01 [−.27, .29]				
10. Cognitive engagement	−.14 [−.40, .14]	−.14 [−.41, .14]	.03 [−.25, .30]	−.11 [−.38, .18]	.30* [.03, .54]	−.26 [−.50, .02]	.31* [.03, .54]	−.12 [−.38, .17]	.60** [.39, .75]			
11. Situational interest	.05 [−.23, .33]	−.27 [−.51, .01]	−.06 [−.33, .22]	−.11 [−.37, .18]	.51** [.27, .69]	−.32* [−.55, −.05]	.11 [−.18, .37]	−.05 [−.32, .23]	.49** [.24, .67]	.60** [.39, .75]		
12. Involvement	−.11 [−.38, .17]	−.06 [−.34, .22]	.14 [−.15, .40]	−.23 [−.47, .06]	.30* [.02, .53]	−.28 [−.52, .00]	.35* [.07, .57]	−.17 [−.42, .12]	.62** [.42, .77]	.76** [.61, .86]	.68** [.50, .81]	

Table 2 (continued)

Variable	1	2	3	4	5	6	7	8	9	10	11	12
13. Knowledge test	.09 [−.20, .36]	−.07 [−.34, .21]	−.23 [−.48, .05]	−.19 [−.45, .09]	.20 [−.08, .45]	−.03 [−.30, .25]	−.08 [−.35, .21]	.33* [.06, .56]	.30* [.03, .54]	.12 [−.16, .39]	.42** [.16, .62]	.21 [−.07, .46]

To better understand the relations between individual prerequisites, learning activities, and learning outcomes, we calculated correlations across all variables. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). Abitur grade: lower values indicate better results according to the German grading system

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

Table 3 Prediction of knowledge test results (N = 52)

	Model 1			Model 2			Model 3		
	b	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Manual rating	1.08	0.49	.032	0.92	0.49	.067	1.00	0.48	.042
Abitur grade				−0.60	0.29	.043	−0.50	0.30	.099
School type				−0.40	0.28	.159	−0.47	0.27	.087
Cognitive abilities							0.08	0.04	.068
Dispositional interest							0.43	0.23	.066
Self-concept							0.38	0.26	.160
Self-control capacity							−0.28	0.21	.189
R^2	.092			.184			.342		
F	4.88*			3.46*			3.12**		

Abitur grade: lower values indicate better results according to the German grading system

*$p < .05$; **$p < .01$; ***$p < .001$

The correlations between the different models for estimating the manual ratings and students' self-reported learning activities and outcomes revealed that the best models were those in which head pose and gaze features were combined with neighbor synchrony ($r = .08$, .43, .39, and .26 for the knowledge test, involvement, cognitive engagement, and situational interest, respectively; Table 5). We calculated the mean correlation (based on Fisher's *z*-transformed correlations) of the three manual annotations (average $r = .74$) and the mean correlation of each rater and the scores from a model combining head pose, gaze features, and neighbor synchrony (average $r = .64$) for the subsample.

Because the model in which head pose and gaze were combined with neighbor's synchrony had the highest correlation with the manual rating, we calculated a linear regression to predict the posttest variables (Table 6). In order to understand the contribution of neighbor's synchrony, we trained our regression models using the same features with and without synchronization information. Adding neighbor's synchrony improved the prediction of all posttest variables and explained at least 2% more variance. However, the manual rating remained superior.

Table 4 Performance of different modalities in engagement in learning estimation depicted as mean squared error (MSE) for regression and Pearson correlations between manual ratings and our models' estimation ($N = 30$)

Modalities	MSE	*r*	*p*
Single students			
Head pose	0.057	.29	.126
Gaze	0.055	.44	.015
Facial expressions	0.056	.44	.014
Head pose + gaze	0.052	.61	.000
3-Combined	0.051	.61	.000
Single students + cosine similarity			
Head pose + gaze (sync)	0.029	.71	.000
3-Combined (sync)	0.050	.70	.000

Table 5 Pearson correlations of different modalities in engagement in learning estimations with post-test variables ($N = 30$)

Modalities	Knowledge test		Involvement		Cognitive engagement		Situational interest	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Single students								
Manual ratings	.14	.468	.64	.000	.62	.001	.53	.003
Head pose	− .17	.392	.05	.799	.02	.914	− .02	.913
Gaze	.11	.582	.19	.335	.16	.414	.23	.236
Facial expressions	− .09	.667	.37	.053	.23	.249	.30	.116
Head pose + gaze	− .03	.867	.41	.029	.37	.053	.21	.286
3-Combined	− .04	.827	.43	.023	.37	.055	.21	.277
Single students + similarity								
Head pose + gaze (sync)	.08	.704	.43	.023	.39	.040	.26	.175
3-Combined (sync)	− .01	.968	.45	.016	.38	.043	.26	.189

Discussion

The present study reported key initial results from the development of a machine vision-based approach for assessing (dis)engagement in the classroom. We were able to find empirical support for the validity of our newly developed manual rating instrument. Furthermore, the machine-learning approach proved to be effective, as shown by its correlation with the manual annotations as well as its ability to predict self-reported learning activities. Finally, as expected, including an indicator of synchrony in the automated analyses further improved its predictive power. Next, we discuss our main results in more detail before turning to the limitations of the present study and the crucial next steps.

Empirical Support for the Newly Developed Approach

The manual rating of visible indicators for (dis)engagement in learning predicted achievement on a knowledge test following a university seminar session. This prediction was robust when we controlled for individual characteristics (research question 1). In terms of validity, self-reported cognitive engagement, involvement, and situational interest were strongly correlated with the manual rating. As these self-reported learning activities reflect students' cognitive processes during the seminar session, we concluded that our manual ratings capture visible indicators that are actually related to (dis)engagement in learning. Therefore, we inferred that it is reasonable to use these manual ratings as a ground truth for our machine vision-based approach.

In the automated analyses of engagement, we used several visible features (head pose, gaze, facial expressions). More specifically, we compared their contribution with visible indicators of (dis)engagement in learning separately and in combination. Our results showed that facial expressions were more strongly correlated with the manual rating than head pose or gaze alone; however, combining the engagement-related features and combining all three visible indicators improved the correlation with the manual annotations substantially, thus emphasizing the complexity of human rating processes. However, we were not able to replicate the prediction of the knowledge test scores by considering these visible features alone (research question 2).

Table 6 Prediction of post-test variables by fused head pose and gaze estimation, fused head pose and gaze estimation plus cosine similarity, and manual rating in subsample (N = 30)

	Estimated rating (head pose + gaze)					Estimated rating (head pose + gaze) + sync					Manual rating				
	b	*SE*	*p*	R^2	*F*	*b*	*SE*	*p*	R^2	*F*	*b*	*SE*	*p*	R^2	*F*
Knowledge test	1.37	8.09	.867	.001	0.03	1.14	2.98	.704	.006	0.15	0.63	0.86	.468	.020	0.54
Cognitive engagement	7.74	3.82	.053	.136	4.10	3.03	1.40	.040	.152	4.67*	1.38	0.35	.000	.380	15.91***
Involvement	13.94	6.05	.030	.170	5.31*	5.37	2.42	.023	.184	5.87*	2.33	0.54	.000	.414	18.34***
Situational interest	5.64	5.17	.286	.044	1.19	2.63	1.88	.175	.070	1.95	1.54	0.48	.003	.286	10.42**

*$p < .05$; **$p < .01$; ***$p < .001$

We expected that additional information concerning interaction with peers and similar behavioral aspects would improve the estimated model. Indeed, adding synchrony by considering the engagement patterns of students' neighbors improved the correlations with the manual rating as well as the prediction of the posttest variables (research question 3). In line with Raca et al.'s (2013) correlative results, our findings indicated that considering neighbor synchrony leads to a better understanding of engagement in predictive models. However, the manual ratings were still better at predicting the knowledge test results as well as self-reported cognitive engagement, involvement, and situational interest. Yet, the similarity between the three different manual raters ($r = .74$) differed from the similarity between the manual annotations and the machine-learning approach ($r = .64$). This difference obviously leaves some room for improvement; however, the approximation that was based on visual parameters and the synchrony with a neighbor's behavior appears to provide reliable results. This raises the question of whether human annotators should also include more than just a single person in their ratings and (unconsciously) consider additional information.

Possible Contributions of an Automated Approach for Assessing Engagement

Our machine-learning approach provides a promising starting point for reducing the effort involved in manual video inspection and annotation, which in turn would facilitate the analysis of larger numbers of individuals and longer videotaped lessons. In addition, such approaches enable the consideration of more complex information on synchronization across students in a way that goes beyond the ability of human observers. This approach is potentially fruitful for both research and practice.

Information from automated analyses of engagement can be used to provide feedback to teachers and improve their skills in monitoring and identifying relevant cues for students' attention in complex classroom interactions. When teachers can notice and identify a lack of engagement, they have the opportunity to adapt their teaching method accordingly and to encourage the students to deal with the learning content actively. Furthermore, by noticing and identifying distracting behavior, teachers get the chance to react to disruptions and ensure the effective use of instruction time. An automated analysis of videos can support novice teachers in developing professional vision skills, and it can provide feedback to teachers in general about the deep structure of their teaching. By making work with videos less effortful, this method could allow videos to be implemented in teacher training more systematically.

Moreover, the annotation of (dis)engagement in learning over time opens up new opportunities for further investigations of classroom instruction by adding a temporal component. This method allows for the detection of crucial events that accompany similar engagement-related behavior across students and provides deeper insights into different effect mechanisms during instruction. Furthermore, this approach can be combined with additional measures. For example, tracking human raters' eye movements can provide insights into where they retrieve their information and what kinds of visible indicators they actually consider. This knowledge can further improve machine vision-based approaches by including the corresponding features. In addition, combining valid visible indicators of students' (dis)engagement in learning with eye-tracking data for the teacher, for example, makes it possible to analyze in more detail what kind of visible indicators attract novice teachers' attention (e.g., Sümer et al. 2018). This information can then be reported back in teacher training to support professional vision skills.

Challenges and Limitations

Our study has several notable limitations that need to be addressed in future research. First, face recognition was not possible for all students due to the occlusion of their faces some or most of the time. For this reason, we had to reduce the sample size for the automated analysis, which in turn reduced the statistical power. Limited data was also an issue in the study by Whitehill et al. (2014), who only found moderate correlations between engagement and learning for this reason. It can thus be assumed that increasing the number of participants recognized by face detection would further improve the linear regression models used to predict self-reported learning activities and learning outcomes. The use of mobile eye trackers for each student is an example of one solution that can provide data for individual students. However, the use of eye trackers is expensive, and when used with children who might touch the glasses too often, it deteriorates the gaze calibration and results in an erroneous analysis of attention. Besides, mobile eye trackers can affect the natural behavior of students, whereas field cameras are pervasive and do not create a significant intervention. To overcome the issue of students being occluded, different camera angles could be helpful in future studies.

Second, a challenging aspect of engagement estimation in our setting was the highly imbalanced nature of our data. Engagement levels on both outer ends of our rating scale were underrepresented. As a direct consequence of the learning setting (a teacher-focused session on statistics), few participants displayed active on-task behavior (e.g., explaining content to others); even less data were collected for visible indicators of disengagement in learning indicating active off-task behavior (e.g., walking around with the intention to interrupt). This imbalance has negative implications for the training of algorithms because greater variability in behavior typically leads to more accurate automated analyses. Whereas human raters are familiar with high levels of variance in an audience's on-task and off-task behavior and use this implicit knowledge in their annotation, the algorithms were trained using only the available data from our three sessions. However, this limitation can be overcome by recording real classroom situations, which will be part of our future work. Although it is not possible to control the intensity of students' (dis)engagement in learning in natural classroom settings, completing more recording sessions and including more participants will eventually lead to a wider distribution of characteristics.

Third, additional research is necessary to validate our approach in schools due to the different target population. This is particularly important because high school students might exhibit a more diverse set of visible indicators of (dis)engagement in learning.

Conclusion

Remote approaches from the field of computer vision have the potential to support research and teacher training. For this to be achieved, valid visible indicators of students' (dis)engagement in learning are needed. The present study provides a promising contribution in this direction and offers a valid starting point for further research in this area.

Funding Information Patricia Goldberg and Ömer Sümer are doctoral students at the LEAD Graduate School & Research Network (GSC1028), funded by the Excellence Initiative of the German federal and state governments. The current research was funded as part of the Leibniz-WissenschaftsCampus "Cognitive Interfaces" by a grant to Ulrich Trautwein, Peter Gerjets, and Enkelejda Kasneci.

Compliance with Ethical Standards

The ethics committee from the Leibniz-Institut für Wissensmedien in Tübingen approved our study procedures (approval #2018-017), and all participants gave written consent to be videotaped.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students "game the system". *Conference on Human Factors in Computing Systems*, 383–390.

Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). *OpenFace 2.0: facial behavior analysis toolkit.* Paper presented at the 13th IEEE International Conference on Automatic Face & Gesture Recognition, ake Placid, NY. https://doi.org/10.1109/WACV.2016.7477553.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*, 463–482.

Bertrams, A., & Dickhäuser, O. (2009). Messung dispositioneller Selbstkontroll-Kapazität. *Diagnostica, 55*(1), 2–10. https://doi.org/10.1026/0012-1924.55.1.2.

Bosch, N., D'Mello, S. K., Baker, R. S., Ocumpaugh, J., Shute, V., Ventura, M., . . . Zhao, W. (2016a). *Detecting student emotions in computer-enabled classrooms.* The Twenty-Fifth International Joint Conference on Artificial Intelligence, 4125-4129.

Bosch, N., D'Mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016b). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems, 6*(2), 1–26. https://doi.org/10.1145/2946837.

Brünken, R., & Seufert, T. (2006). Aufmerksamkeit, Lernen, Lernstrategien. In H. Mandl & H. F. Friedrich (Eds.), *Handbuch Lernstrategien* (pp. 27–37). Göttingen: Hogrefe.

Büttner, G., & Schmidt-Atzert, L. (2004). *Diagnostik von Konzentration und Aufmerksamkeit*. Hogrefe Verlag.

Carroll, J. B. (1963). A model for school learning. *Teachers College Record, 64*(8), 723–733.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823.

Connell, J. P. (1990). Context, self, and action: a motivational analysis of self-system processes across the life span. In D. Cicchetti & M. Beeghly (Eds.), *The self in transition: infancy to childhood* (Vol. 8, pp. 61–97). Chicago: The University of Chicago Press.

Cortina, K. S., Miller, K. F., McKenzie, R., & Epstein, E. (2015). Where low and high inference data converge: validation of CLASS assessment of mathematics instruction using mobile eye tracking with experts and novice teachers. *International Journal of Science and Mathematics Education., 13*(2), 389–403. https://doi.org/10.1007/s10763-014-9610-5.

Cumming, G. (2014). The New Statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology, 92*(1), 53–78. https://doi.org/10.1348/000712601162103.

Ehrhardt, K. J., Findeisen, P., Marinello, G., & Reinartz-Wenzel, H. (1981). Systematische Verhaltensbeobachtung von Aufmerksamkeit im Unterricht: Zur Prüfung von Objektivität und Zuverlässigkeit. *Diagnostica*, 281–294.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: manual*. Palo Alto: Consulting Psychologists Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Erickson, F. (2007). Ways of seeing video: toward a phenomenology of viewing minimally edited footage. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video research in the learning sciences* (2nd ed., pp. 145–155). New York: Routledge.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research, 9*(Aug), 1871–1874.

Frank, B. (2014). *Presence messen in laborbasierter Forschung mit Mikrowelten: Entwicklung und erste Validierung eines Fragebogens zur Messung von Presence* .Wiesbaden: Springer-Verlag.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059.

Fujii, K., Marian, P., Clark, D., Okamoto, Y., & Rekimoto, J. (2018). Sync class: visualization system for in-class student synchronization. Paper presented at the Proceedings of the 9th Augmented Human International Conference, New York, NY.

Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology, 48*, 67–84. https://doi.org/10.1016/j.cedpsych.2016.09.003.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience, 8*(385). https://doi.org/10.3389/fnins.2014.00385.

Girard, J. M. (2014). CARMA: software for continuous affect rating and media annotation. *Journal of Open Research Software, 2*(1), e5.

Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. (2013). *Automatically recognizing facial expression: predicting engagement and frustration.* Paper presented at the Educational Data Mining 2013, Memphis, TN.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34.

Heller, K., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision: KFT 4–12+ R* [Cognitive abilities test for 4th to 12th grade students]. In: Weinheim: Beltz-Test.

Helmke, A., & Renkl, A. (1992). Das Münchener Aufmerksamkeitsinventar (MAI): Ein Instrument zur systematischen Verhaltensbeobachtung der Schüleraufmerksamkeit im Unterricht. *Diagnostica, 38*, 130–141.

Hommel, M. (2012). Aufmerksamkeitstief in Reflexionsphasen- eine Videoanalyse von Planspielunterricht. *Wirtschaft und Erziehung, 1-2*, 12–18.

Karweit, N., & Slavin, R. E. (1981). Measurement and modeling choices in studies of time and learning. *American Educational Research Journal, 18*(2), 157–171. https://doi.org/10.3102/00028312018002157.

Kleinknecht, M., & Gröschner, A. (2016). Fostering preservice teachers' noticing with structured video feedback: results of an online- and video-based intervention study. *Teacher and Teacher Education, 59*, 45–56. https://doi.org/10.1016/j.tate.2016.05.020.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe 1: "Aufgabenkultur" und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Eds.), *TIMSS. Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Videodokumente* (pp. 33–49). Berlin: Max-Planck-Institut für Bildungsforschung.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". In M. Prenzel & L. Allolio-Naecke (Eds.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (pp. 127–146). Waxmann: Münster.

Knogler, M., Harackiewicz, J. M., Gegenfurtner, A., & Lewalter, D. (2015). How situational is situational interest? Investigating the longitudinal structure of situational interest. *Contemporary Educational Psychology, 43*, 39–50. https://doi.org/10.1016/j.cedpsych.2015.08.004.

Krumpe, T., Scharinger, C., Rosenstiel, W., Gerjets, P., & Spueler, M. (2018). Unity and diversity in working memory load: Evidence for the separability of the executive functions updating and inhibition using machine learning. *Biological Psychology, 139*, 163-172. https://doi.org/10.1016/j.biopsycho.2018.09.008.

Lachner, A., Jarodzka, H., & Nückles, M. (2016). What makes an expert teacher? Investigating teachers' professional vision and discourse abilities. *Instructional Science, 44*(3), 197–203. https://doi.org/10.1007/s11251-016-9376-y.

Lahaderne, H. M. (1968). Attitudinal and intellectual correlates of attention: a study of four sixth-grade classrooms. *Journal of Educational Psychology, 59*(5), 320–324. https://doi.org/10.1037/h0026223.

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: a meta-analysis. *Social Behavior and Personality, 46*(3), 517–528. https://doi.org/10.2224/sbp.7054.

Lizdek, I., Sadler, P., Woody, E., Ethier, N., & Malet, G. (2012). Capturing the stream of behavior: a computer-joystick method for coding interpersonal behavior continuously over time. *Social Science Computer Review, 30*(4), 513–521. https://doi.org/10.1177/0894439312436487.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality, 74*(2), 403–456. 00380. https://doi.org/10.1111/j.1467-6494.2005.00380.x.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59*(1), 14–19. https://doi.org/10.1037/0003-066X.59.1.14.

McKinney, J. D., Mason, J., Perkerson, K., & Clifford, M. (1975). Relationship between classroom behavior and academic achievement. *Journal of Educational Psychology, 67*(2), 198. https://doi.org/10.1037/h0077012.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734.

Pauli, C., & Lipowsky, F. (2007). Mitmachen oder Zuhören? Mündliche Schülerinnen- und Schülerbeteiligung im Mathematikunterricht. *Unterrichtswissenschaft, 35*(2), 101–124.

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

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33–40. https://doi.org/10.1037/0022-0663.82.1.33.

Poh, M., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 57*(5), 1243–1252.

Posner, M. I. (1988). Structures and functions of selective attention. In T. Boll & B. Bryant (Eds.), *Master lectures in clinical neuropsychology* (pp. 173–202). Washington, D.C.: American Pschology Association.

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

Raca, M. (2015). Camera-based estimation of student's attention in class. Retrieved from https://infoscience.epfl.ch/record/212929.

Raca, M., & Dillenbourg, P. (2013). *System for assessing classroom attention.* Paper presented at the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium.

Raca, M., Tormey, R., & Dillenbourg, P. (2013). *Student motion and it's potential as a classroom performance metric.* Paper presented at the 3rd International Workshop on Teaching Analytics (IWTA), Paphos, Cyprus.

Raca, M., Tormey, R., & Dillenbourg, P. (2014). *Sleepers' lag-study on motion and attention.* Paper presented at the Fourth International Conference on Learning Analytics And Knowledge, Indianapolis, IN.

Raca, M., Kidzinski, L., & Dillenbourg, P. (2015). *Translating head motion into attention-towards processing of student's body-language.* Paper presented at the 8th International Conference on Educational Data Mining, Madrid, Spain.

Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A. A., & Curby, T. W. (2015). To what extent do teacher-student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning? *Journal of Educational Psychology, 107*(1), 170–185. https://doi.org/10.1037/a0037252.

Sherin, M. G., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education, 60*(1), 20–37. https://doi.org/10.1177/0022487108328155.

Star, J. R., & Strickland, S. K. (2008). Learning to observe: using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education, 11*(2), 107–125. https://doi.org/10.1007/s10857-007-9063-7.

Stipek, D. (2002). Good instruction is motivating. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 309–332). San Diego: Academic Press.

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

Sümer, Ö., Goldberg, P., Stürmer, K., Seidel, T., Gerjets, P., Trautwein, U., & Kasneci, E. (2018). Teachers' perception in the classroom . Paper presented at the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT.

Trautwein, U., Gerjets, P., & Kasneci, E. (2017). *A cognitive interface for educational improvement: assessing students' attentional focus in the classroom* .University of Tuebingen.

Whitehill, J., Serpell, Z., Lin, Y., Foster, A., & Movellan, J. R. (2014). The faces of engagement: automatic recognition of student engagement from facia expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(1), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163.

Wolff, C. E., Jarodzka, H., van den Bogert, N., & Boshuizen, H. P. A. (2016). Teacher vision: expert and novice teachers' perception of problematic classroom management scenes. *Instructional Science, 44*(3), 243–265. https://doi.org/10.1007/s11251-016-9367-z.

Yamamoto, T., & Imai-Matsumura, K. (2012). Teachers' gazes and awareness of students' behavior: unsing an eye tracker. *Innovative Teaching, 2*(6). https://doi.org/10.2466/01.IT.2.6.

Yoshida, R., Nakayama, T., Ogitsu, T., Takemura, H., Mizoguchi, H., Yamaguchi, E., . . . Kusunoki, F. (2014). Feasibility study on estimating visual attention using electrodermal activity. *8th International Conference on Sensing Technology.*

Zaletelj, J. (2017). *Estimation of students' attention in the classroom from kinect features.* Paper presented at the 10th International Symposium on Image and Signal Processing and Analysis (ISPA), Ljubljana, Slovenia.

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing, 2017*(1), 80–12. https://doi.org/10.1186/s13640-017-0228-8.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017). S3FD: *single shot scale-invariant face detector.* Paper presented at the The IEEE International Conference on Computer Vision (ICCV), Venice, Italy.

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

Affiliations

Patricia Goldberg[1] · Ömer Sümer[1] · Kathleen Stürmer[1] · Wolfgang Wagner[1] · Richard Göllner[1] · Peter Gerjets[2] · Enkelejda Kasneci[3] · Ulrich Trautwein[1]

[1] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastr. 6, 72072 Tubingen, Germany

[2] Leibniz-Institut für Wissensmedien, Schleichstraße 6, 72076 Tübingen, Germany

[3] Chair of Media Informatics and Human-Computer Interaction, University of Tübingen, Sand 13, 72076 Tubingen, Germany