Zeitschrift für Erziehungswissenschaft (2023) 26:677–702
https://doi.org/10.1007/s11618-023-01166-8

SCHWERPUNKT

Fast-and-frugal means to assess reflection-related reasoning processes in teacher training—Development and evaluation of a scalable machine learning-based metric

Lukas Mientus · Peter Wulff · Anna Nowak · Andreas Borowski

Received: 2 February 2022 / Revised: 22 March 2023 / Accepted: 20 April 2023 / Published online: 6 July 2023

Abstract Reflection is hypothesized to be a key component for teachers' professional development and is often assessed and facilitated through written reflections in university-based teacher education. Empirical research shows that reflection-related competencies are domain-dependent and multi-faceted. However, assessing reflections is complex. Given this complexity, novel methodological tools such as non-linear, algorithmic models can help explore unseen relationships and better determine quality correlates for written reflections. Consequently, this study utilized machine learning methods to explore quality correlates for written reflections in physics on a standardized teaching situation. $N = 110$ pre- and in-service physics teachers were instructed to reflect upon a standardized teaching situation in physics displayed in a video vignette. The teachers' written reflections were analyzed with a machine learning model which classified sentences in the written reflections according to elements in a reflection-supporting model. A quality indicator called *level of structure* (LOS) was devised and further used to validate machine learning classifications against experts' judgements. Analyses show that LOS is positively correlated with experts' judgements on reflection quality. We conclude that LOS of

✉ Lukas Mientus · Anna Nowak · Prof. Dr. Andreas Borowski
Physics Education Research, University of Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam-Golm, Germany
E-Mail: lukas.mientus@uni-potsdam.de

Anna Nowak
E-Mail: anna.nowak@uni-potsdam.de

Prof. Dr. Andreas Borowski
E-Mail: adreas.borowski@uni-potsdam.de

Jun.-Prof. Dr. Peter Wulff
Physics and Physics Education Research, Heidelberg University of Education, Postfach 10 42 40, 69032 Heidelberg, Germany
E-Mail: peter.wulff@ph-heidelberg.de

a written reflection is one important indicator for high-quality written reflections which is able to exclude typical quality correlates such as text length. With the help of the machine learning model, LOS can be useful to assess pre-service physics teachers written reflections.

Keywords Reflection · Quality · Assessment · Machine learning

Effektives Bewerten reflexionsbezogener Argumentationsprozesse in der Lehrkräftebildung – Entwicklung und Evaluation einer skalierbaren Metrik mittels maschinellen Lernens

Zusammenfassung Reflexion gilt als Schlüsselkomponente für die berufliche Entwicklung von Lehrkräften und wird in der universitären Lehrkräftebildung häufig durch schriftliche Reflexionen bewertet und gefördert. Die empirische Forschung zeigt, dass reflexionsbezogene Kompetenzen bereichsabhängig und vielschichtig sind. Die Bewertung von Reflexionen ist jedoch komplex. Angesichts dieser Komplexität können neuartige methodische Instrumente wie nichtlineare, algorithmische Modelle dazu beitragen, hintergründige Beziehungen zu erforschen und Qualitätskorrelate für schriftliche Reflexionen besser zu bestimmen. In dieser Studie wurden daher Methoden des maschinellen Lernens eingesetzt, um Qualitätskorrelate für schriftliche Reflexionen in Physik in einer standardisierten Unterrichtssituation zu untersuchen. $N = 110$ (angehende) Physiklehrkräfte verfassten im Rahmen einer Videovignette eine schriftliche Reflexion über eine standardisierte Unterrichtssituation in Physik. Die schriftlichen Reflexionen der Physiklehrkräfte wurden mit einem maschinellen Lernmodell analysiert, das die Sätze in den schriftlichen Reflexionen nach den Elementen eines Rahmenmodells für Reflexion klassifizierte. Auf Basis dieses Algorithmus wurde ein Qualitätsindikator mit der Bezeichnung *Level of Structure* (LOS) entwickelt und gegenüber Einschätzungen von Experten validiert. Die Analysen zeigen, dass LOS positiv mit den Einschätzungen der Experten zur Qualität der Reflexion korreliert. Wir kommen zu dem Schluss, dass unser LOS einer schriftlichen Reflexion ein wichtiger Indikator für qualitativ hochwertige schriftliche Reflexionen ist, der typische Qualitätskorrelate wie die Textlänge ausschließen kann. Mit Hilfe des maschinellen Lernmodells kann ein LOS nützlich sein, um die schriftlichen Reflexionen von angehenden Physiklehrern zu bewerten.

Schlüsselwörter Reflexion · Qualität · Bewertung · Machine Learning

The professional knowledge base of teachers was determined to be among the most important factors that predict effective teaching in the STEM (science, technology, engineering, mathematics) fields (Kunter et al. 2011; Sadler et al. 2013). For the professional knowledge of pre-service teachers in STEM fields, action-oriented professional knowledge purportedly develops through planning and enacting lessons and reflecting upon them (Carlson et al. 2019). In teacher training, however, enacting own lessons is mostly reserved for later stages. Simulations of practice, such as video-taped lessons, have been found to be a valuable means to support the

development of action-oriented professional knowledge in complexity-reduced environments (van Es and Sherin 2008). Reflection is oftentimes utilized as a means to elicit professional knowledge and apply it in reasoning processes on authentic teaching situations, and thus facilitate pre-service STEM teachers to become reflective practitioners (Hume 2009; Schön 1983). As such, reflection-related reasoning processes are mostly elicited with open-ended response formats such as reflective diaries, reports, or protocols.

However, either assessing these open-ended response formats where reflection-related reasoning processes are elicited, and providing guidance on the basis of these responses is riddled with challenges. For once, reliably coding spoken or written reflections of pre-service and in-service teachers could only be achieved with tremendous expense of resources, e.g., multiple rounds of coding where convergence of interpretation of the data is slow (Abels 2011; Kember et al. 1999). Human interrater agreement was oftentimes low, because evidence for reflection-related reasoning processes that are represented in the form of natural language is inherently ambiguous (Jurafsky 2003; Sparks-Langer et al. 1990). Once agreement in a specific research project on the interpretation of reflection-related data was achieved, it was argued that sharing the coding rubrics with other researchers was largely unfeasible. Biernacki (2014) contends that "the classifying demanded by a coding operation is so delicate that its validity is perhaps too tentative for others to build on" (p. 177). It would be important to examine to what extent reflection-related processes across contexts are similar to each other and how they can be facilitated through targeted guidance.

The recent resurgence of computer-based analysis tools, in particular artificial intelligence (AI)-based research methods, provide novel potentials to assess complex constructs such as reflection-related reasoning processes in an evidence-centered way. Especially the natural language processing (NLP) sought tremendous progress with the advent of large language models that are pretrained on large datasets and can be fine-tuned in specific research contexts (e.g. Wulff et al. 2023). This novel paradigm in research was called data-driven discovery (Hey et al. 2009). Analyses of reflection-related reasoning processes in general educational settings and in discipline-specific settings yielded some insights on applicability of AI-based methods for assessing reflection-related reasoning processes. Mostly, AI methods can facilitate reliable, automated coding, and assessment of themes that are addressed in a written reflection (Ullmann 2019). What is unclear today, however, is in what ways the outputs of AI-based methods (e.g., automatically generated classifications) can be further used to estimate the quality of a written reflection.

In this study we examine the extent to which automatically generated classifications for pre-service teachers written reflections correlate with expert ratings. We situate our study in the context of discipline-based education (here: physics), as we expect the reflection-related reasoning to show most variance with respect to knowledge-related factors, rather than more general educational topics such as cognitive activation, classroom management, and constructive feedback. After all, expertise in teaching and beyond is "specific to a domain" (Berliner 2001, p. 463). Our findings have implications for designing intelligent tutoring systems for reflection-related reasoning processes. Today, guidance can be easily provided through

chatbots in generative language models (e.g., ChatGPT). For education research it is important to reconcile AI-generated model outputs with actual human ratings to assure alignment of human and machine (Christian 2021).

1 Reflection in teacher training programs

Reflection on teaching lessons is regarded as a key component for teachers' professional development (Christof et al. 2018; Korthagen and Kessels 1999; Sorge et al. 2018). Consequently, many effective teacher training programs include "reflection" (Darling-Hammond 2012). Researchers have recurrently pointed out problems of finding a clear-cut, agreed-upon definition of what reflection actually is (Häcker 2022; Leonhard 2022). As a working definition in this study, we contend that to learn reflective skills, teachers are meant to go through a process of structured analysis in which a relationship is established between teacher's own knowledge, skills, attitudes/beliefs and/or dispositions and teacher's own situation-specific thinking and behavior [...] with the aim of (further) developing teacher's own knowledge, attitudes ... and/or teacher's own thinking and behavior (von Aufschnaiter et al. 2019). Reflection is thus not a unitary construct, but rather comprised of many different aspects that interact in complex ways with each other.

Reflection-related reasoning processes are typically elicited and examined through standardized teaching situations and prompts for reflection (Kori et al. 2014). While video-recoding of one's own teaching situation has the benefit to be more intense, standardization enables researchers to better control for the specific problems that are relevant in the situation, and for comparison of different reflection-related reasoning processes among pre-service teachers. Standardized teaching situations usually show a problematic situation that can be very subject specific. Simultaneously, it seems to be irrelevant for the noticing of problems whether the video recording is of one's own situation or that of another teacher (Seidel et al. 2011). To guide reflection-related reasoning processes, reflection prompts, e.g., in the form of reflection-supporting models are often provided to the pre-service teachers (Kori et al. 2014). Research on noticing and professional vision focuses more on the particular information-processing in the situation, and the interpretations (Cappell 2013). Reflection research is more encompassing, as the pre-service teachers are required to explicitly relate their own prior knowledge and attitudes to the situation, and derive consequences for their personal professional development (von Aufschnaiter et al. 2019). Both foci, noticing/vision and reflection are important, though they have a different scope and a different function for the professional development.

Reflection-related reasoning processes can be elicited through open-ended, natural language-based assessment formats such as constructed responses. Constructed response items allow teachers to use their own knowledge and reasoning to reflect a teaching situation (Poldner et al. 2014). Content analysis is then used to analyze the written products (Hume 2009). On this basis, researchers found that pre-service teachers oftentimes include imprecise, positive evaluations on teaching situations (Mena-Marcos et al. 2013). Moreover, it was difficult for teachers to transcend the personal realm of knowledge (Ovens and Tinning 2009). Without scaffolding, rarely

do teachers reach a level of critical reflection where personal assumptions and learning-relevant processes are analyzed in sufficient depth (Abels 2011; Leonhard and Rihm 2011). Kost (2019) found that (at least for physics teachers) minimal guidance on what to include in a reflection-related reasoning process was effective to motivate the pre-service teachers to more encompassingly reflect on the teaching situations. Researchers consequently argued that reflection-related reasoning processes should include information on situation and circumstances of teaching, and an accurate and precise description (enabling the distancing from experience) of the teaching situation (form a basis for reasoning) (Hatton and Smith 1995; Onyx and Small 2001). For reflection-related reasoning processes to facilitate professional growth and development of applicable professional knowledge, judging and evaluating a situation, and thinking about it from alternative perspectives is also important (Aeppli and Lötscher 2016; Krieg and Kreis 2014). Moreover, devising consequences should focus attention of pre-service teachers on their own professional development (von Aufschnaiter et al. 2019). What is a higher-quality reflection is not up to us, cause the discourse around reflection, reflection related competencies and quality estimation is divers (Mientus et al. in press). For this study we just want to point out that a higher-quality reflection might be more discursive and less descriptive (Abels 2011; Hatton and Smith 1995; Mena-Marcos and Tillema 2006).

2 Facilitating reflective writing analytics with AI-methods

While qualitative studies, based on manual content analysis, provides important insights into the constitution of written reflections, there arise particular challenges that make this research program difficult to extend. For example, Leonhard and Rihm (2011) report that they cannot scale coding of written reflections because it is resource consuming. Moreover, many researchers engaged in content analysis of written reflections noted difficulties in reaching interrater agreement, eventually caused by unclear coding units and high-inferential categories (e.g. Kost 2019). Uncertainty in coding units and categories might cause the interrater agreement to decrease (Kuckartz 2022; Mayring 2022), which then makes communicative validation necessary. While rather short constructed responses mostly lead to acceptable interrater agreement, longer written reflections (as encountered in practice) that frequently comprise hundreds of words are much more difficult to handle. Longer texts refer to complex systems where words are embedded in sentences, sentences are embedded in paragraphs, and paragraphs are embedded within the entire text.

Data-driven and computer-based approaches to reflective writing analytics can offer novel potentials, because they are (in principle) not constrained by human frailties such as limited working-memory, varying linguistic competence, and differences in prior knowledge. In practice, however, AI-methods have many similar problems as human raters, and are currently also constrained by the knowledge base they can tap into. Even though, worth mentioning, inductive information and knowledge extraction from textual data (including the entire Wikipedia or the Internet) reached novel heights with generative language models such as GPT4. Aforementioned passed

medical, educational, or mathematics exams on par with human achievement, however, without any specific training (Katz et al. 2023).

ML and NLP have been used for a wide range of applications, for example assessment of constructed, open-ended responses and even feedback generation (Shermis et al. 2019; Zhai et al. 2020). ML enables computer programs to learn from new data in order to improve their performance in a given task. NLP on the other hand refers to the effective representations and automated processing of natural language (Mitchell 1997). ML approaches include, among others, supervised and unsupervised learning. In supervised ML annotated data, e.g., sentences with a relevant code, are presented and the ML algorithm learns this mapping. In unsupervised ML patterns can be explored in data or clusters extracted. For example, in text analysis it is possible to retrieve topics in documents such as essays which can represent themes that the authors address. Supervised ML is particularly fruitful, if a theory or a model exists that allows human raters to annotate the data. This enables automated assessment of unseen samples, which could be used for adaptive feedback.

Wulff et al. (2020) trained a classification model in the context of the reflection-related reasoning model of Nowak et al. (2019). Wulff et al. (2021) improved the supervised ML-model to annotate written reflections and to give feedback for pre-service teachers and their university teachers (Mientus et al. 2021). Using unsupervised ML Wulff et al. (2022) and Wulff et al. (2023) started modelling topics in written reflections to detect reflection-related reasoning processes in more detail. For the assessment of reflection-related reasoning processes, and the scaffolding of such, it is necessary that educational researchers verify the inferences of quality assessments using ML-based methods (critically).

3 ML-based coding in a specific context

In order to continue an existing coding task, new experts always have to be trained. Since ML-based methods can achieve comparable reliability results, Wulff et al. (2020) trained a classification model to analyze reflection-related reasoning in physics teachers' written reflections using a word corpus called BERT. BERT stands for Bidirectional Encoder Representations from Transformers. It is a pre-trained deep learning model that uses a transformer architecture to analyze and understand natural language text. BERT was developed by Google and released in 2018, and it has been trained on a large corpus of text data, allowing it to learn contextual relationships between words in a sentence. BERT has to be trained with sentence-based encodings for a specific task and will then be able to apply these encodings to unfamiliar texts. The basis for the sentence-based classification model of Wulff et al. (2020) is a model for reflection-related reasoning processes, developed by Nowak et al. (2019). Furthermore, the model was developed in the specific context of written reflections in physics-teaching internships in university-based teacher education programs. It focusses on the reflection-related reasoning process and excludes teacher performance. This reflection-supporting model conceptualizes the abovementioned elements as constitutive reflection elements in a written reflection (Nowak et al. 2019). These elements are (1) descriptions of a specific teaching situ-

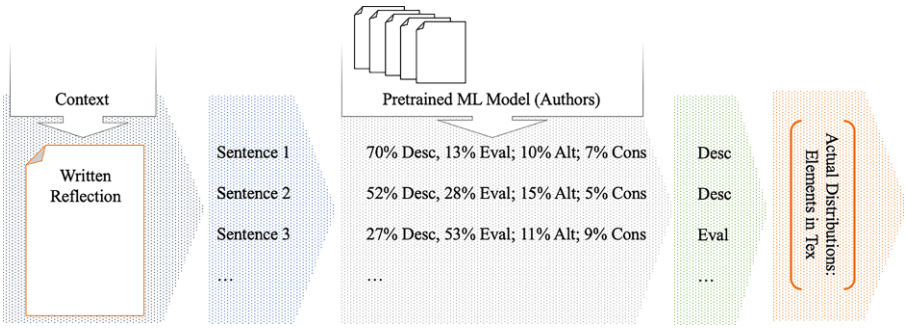

Fig. 1 Scheme of sentence wide classification of a written reflection using pretrained ML model

ation in class, where teachers write about their own and students' actions (including circumstances of the lesson taught, where the reflecting teachers provide details on the setting, class composition, and their set learning goals), (2) evaluations on how the teachers felt during the situation (and why), where teachers judge teachers' and students' performance, (3) alternatives of action for the observed activities, where teachers consider actions that could have done differently to improve the outcomes, and (4) consequences they draw for their own further professional development, where teachers relate their reflections of the situation to their goals for achieving their own professional development.

For the training of the ML model, written reflections were subdivided into sentences which are transformed into features (predictor variables) that will allow the model to predict probabilities for each of the reflection elements for this segment (Wulff et al. 2020). In this way, a most probable reflection element can be assigned to each sentence. Thus, an original information (i.e. a written reflection) can be converted into a dataset that contains only the assigned reflection elements per sentence. This data set provides information about which sentence of the reflection text corresponds to descriptions, evaluations, alternatives, or consequences. The complex text data of a written reflection, which maps the process of reflection-related reasoning, was modelled and the raw data was simplified to an actual distribution of the used reflection elements (see Fig. 1).

In the study of Wulff et al. (2020), "the used words in a segment were anticipated to be an important feature for representing the segments given that Description likely elicits process verbs like 'conduct (an experiment),' Evaluation might be characterized by sentimental words, such as 'good' or 'bad,' and Alternatives requires writing in conditional mode (Ullmann 2019)" (Wulff et al. 2020, p. 7). Thus, the classifier is primarily based on the analysis of words, as this is often used as a starting point for modelling (Ullmann 2019). Nevertheless, it should be noted that the method of Wulff et al. (2020) is not an analysis of word frequency, but the data model BERT forms the basis of all analyses. It is important to understand that after training, a sentence can be assigned to a code by the ML-Model, even if it was assigned to another code for training the ML-Model, because the trained model is based on a large number of coded segments. For this reason, it is important to report a measure of performance using an ML model and relate this to the interrater reliability of human raters.

The ML model was trained and compared in terms of its classification performance through human-machine-agreement. The regularized multinomial logistic regression was found to be the most appropriate classifier for the present purposes, with an F1 score of 0.56 on the test data set (Wulff et al. 2020). Wulff et al. (2020) criticize that about 11% of the human-human comparisons and 40% of the computer-human comparisons are unclear in the areas of circumstances, descriptions and evaluations. It is further pointed out that one reason for the confusion between these elements may be due to inaccuracies in the students' written reflections. As part of the analysis of pre-service physics teachers' written reflections even without ML, Kost (2019) noted that students had difficulty separating descriptive and evaluative texts in the reflections. A deep-learning based approach substantially improved the classification accuracy for this reflection-supporting model and written reflections to an F1 score for the held-out test data to 0.81, which can be considered substantial human-machine-agreement (Wulff et al. 2022). Hence, this model is judged to be a well-versed assessment instrument for written reflections according to the reflection-supporting model. It might help to answer novel research questions that rely on annotated datasets based on the reflection-supporting model.

Studies in reflective writing analytics (Buckingham Shum et al. 2017; Ullmann 2019; Wulff et al. 2023) indicate that computer-based methods such as machine learning are well versed to analytically assess written reflections. So, potentials to advance reflective writing analytics and efficiently model scaffolding for novice teachers' reflections comes from ML research on written reflections. But the question of how to scale quality metric of reflection-related reasoning processes using ML-based methods seems to be unsettled—yet. For this reason, we would like to propose, validate and discuss a metric for fast and frugal means in this paper.

4 Research questions

This study seeks to use a validated ML model in order to develop and evaluate a scalable metric to assess reflection-related reasoning processes in written reflections. In order to ensure that results are not based on a fragile database, the sample will be described using the validated ML method.

RQ 1: To which extent are the sample and the validated ML method representative for written reflections?

Assuming that the written reflections turn out to be predominantly descriptive and that this can be determined with the ML model, we will develop the metric using an expert reflection and evaluate it. In our approach, we determine the quality of a text by the occurrence and order of the elements. The question is whether texts rated higher in quality in this way are actually higher in quality according to expert assessment. The purpose of this study is to verify that.

RQ 2: To what extent do evaluations of experts in observation and feedback of teaching science and estimation based on the ML model yield comparable quality assessments of written reflections?

Does our ML-based metric prove valuable through expert rating, we examine if the extracted categories by the ML models and the metric can help us refine

our understanding of what determines quality in written reflections. Our analyses of the reflection-related reasoning processes focus on reasoning-based features instead of in-depth content analysis to prove and achieve fast-and-frugal means for an assessment of written reflections.

RQ 3: What quality indicators can be automatically extracted from the ML analysis of written reflections and to what extent do they correlate with the developed metric?

5 Method

5.1 Data gathering

As pre-service and novice teachers in particular benefit from scaffolding (Lai and Calandra 2010), pre- and in-service physics teachers are approached for this study. The physics teachers were recruited at various university locations in Germany and at seminar locations in the second phase of teacher training programs. The context of physics teachers was chosen, as physics is considered to be a difficult and knowledge-intensive subject where models to support reflection might be beneficial for novice teachers to appreciate how difficult physics is for students (Hume 2009; Sorge et al. 2018).

5.2 The standardized teaching situation in a video vignette

An online video vignette was implemented to enable pre-service physics teachers to make vicarious experiences. Following the assumption that reflective writing is an important pedagogical practice, which is limited by the available classroom time (Ullmann 2017), reflection on observed or videotaped lessons opens up as a suitable alternative learning opportunity to practice reflective writing in university-based teacher education programs. If pre-service teachers are able to apply their learning in practical simulations, they can concentrate on specific areas of teaching without breaking it down into disconnected, isolated actions (Wyss 2018). Among others, the complex or immediate nature of classroom situations and teachers' reaction to it can be captured in a vignette (Lindmeier 2011). Therefore, a quality assessment by comparing response formats to a vignette with an expert norm seems to be a proven practice (e.g. Oser and Heinzer 2009; Oser et al. 2010). However, it is important to pay attention to heuristics concerning how to choose and how to use a video for a vignette. Blomberg et al. (2013) presents five research-based heuristics in order to create a well-conceptualized learning environment. Our video vignette does not aim at supporting reflective competence but at measuring it. Therefore, we focus only on the third heuristic by choosing an appropriate video material. We selected a suitable segment of a videographed lesson showing authentic classroom situation on the one hand and many observable actions and experiments to reflect on on the other hand.

Participants were instructed to (1) read an introduction in the study (its interest, aims, and conditions), (2) watch the standardized teaching situation, (3) write a reflection, and (4) answer further demographical questions. In the first part, the

participants read an introduction to the reflection model and how the model can help structuring a reflective process. Here, each reflection element is explained and stimulating questions are formulated, which remain readable for the participants throughout the entire process. The participants had the task to (a) describe the observed teaching situation, (b) make justified evaluations, (c) formulate alternatives as well as (d) derive consequences for his/her professional development. After a short text that introduces the circumstances of the teaching situation, the 17-minute video clip follows in a second part, which the participants can watch once. The video shows a physics lesson in a ninth-grade class of a high school. It is a teacher-led introductory lesson on the topic of the free fall. The teaching situation can be described in two steps. During the first step, the teacher conducted various experiments involving falling objects such as two masses of different shape or mass and a screw and a feather inside a vacuum tube. The pupils formulated hypotheses regarding the results of the experiments, such as predictions which of the two objects would touch the ground first. In the second step the teacher defines the concept of free fall and the students discuss the nature of free fall using explanations with the experiments. After watching the videography, the participants write a reflection text in an unlimited text field. The texts form the central data material for the further analyses. Finally, the participants provided some demographical data. All collected data was analyzed with a pseudonym code to comply with data protection requirements.

5.3 Scaling reflective writing metrically

In order to determine to what extent which element might occur, three professors of physics education and research were recruited to provide in-depth analyses of the teaching situation. These experts provided a written report that identified opportunities for pedagogical reasoning in the teaching situation. The experts collaboratively created a written reflection of the teaching situation which also follows the reflection-supporting model of Nowak et al. (2019). The experts were able to watch the video in detail independently of each other (several times or intermittently) and took notes. One expert formulated a detailed written reflection from his own bullet points. Here, all observations were first formulated before building up to justified evaluations and alternative teaching methods. Finally, the consequences that teachers can draw from the observation for their own professionalization were discussed. The two other experts integrated their notes into this written reflection, one after the other, so that a collaborative expert-based reflection emerged. In line with our RQs, only structural aspects of this collaborative written reflection are considered in this study, not statements on content. On the basis of experts' collaborative written reflection, 35% descriptions, 35% evaluations, 15% alternatives, and 15% consequences were chosen as a normative distribution of elements for a written reflection for this study. We posit that this distribution forms a reasonable distribution where no single element is disproportionately represented or missed out. We also assume that the elements appear in blocks and no interchanging of elements appears in high-quality reflections. While this assumption certainly does not cover all high-quality reflections (for example, some students arrange their texts so that they describe-evaluate-reason on a certain problem, and then describe-evaluate-reason on

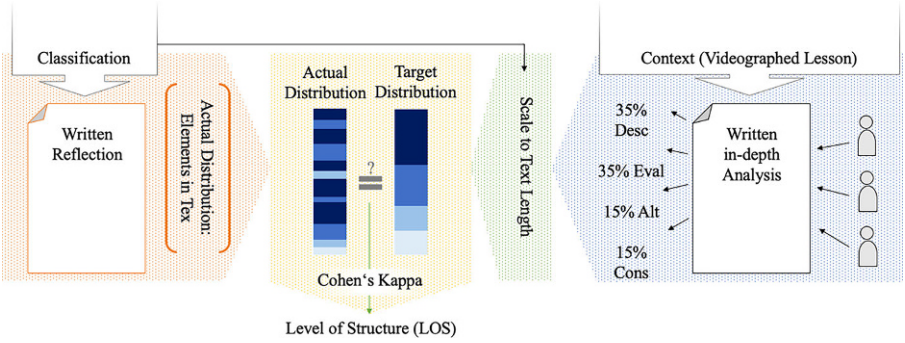

Fig. 2 Scheme of calculating level of structure (*LOS*)

the next problem), we observed that most students arrange their texts in this block-like composition.

To generate an empirically-based quality indicator for teachers' written reflections a comparator between each written reflection and the collaborative written reflection based on the reflection elements and their arrangement has been applied. Based on the text length of a written reflection, a target distribution was determined for each text that would represent a reflection structure according to our normative distribution (35% descriptions, 35% evaluations, 15% alternatives, 15% consequences). In a further step, this target distribution was compared with the reflection structure as determined by the ML model. Here the agreement of the actual distribution of a written reflection and the target distribution was proved and measured as Cohen's kappa. Figure 2 depicts this process in a diagram. In this way, each text was assigned a level of structure (LOS), i.e., the Cohen's kappa score between observed percentages and normative percentages. A LOS of 1 would correspond to a written reflection which has a sequencing of reflection elements according to the reflection-supporting model and matches the normative distribution of the elements. A level of structure of 0 translates into a maximum degree of disorder of the reflection elements in the text; level-1 means a sequencing of the reflection elements in reverse order. The actual distributions in two written reflections may represent different reasoning processes even though the percentages of the four elements (or even the text lengths) may be identical. A different LOS could be a result. Accordingly, texts can have an identical LOS even though their elements are distributed differently in percentage terms. Thanks to the scaling of the expert's reflection by the text length to a target distribution, we hope to have found a structural measure in the LOS that explicitly excludes the text length, which is often associated with the quality of a written reflection (Chodorow and Burstein 2004; Leonhard and Rihm 2011).

5.4 Experts' quality assessment of written reflections

To answer RQ 1, LOS hast to be validated. To do so, we divided the texts into quantiles according to LOS. To clearly separate the extreme groups, we only used the first, third and fifth quantile. Thus, we obtain three separable groups of well-structured (i.e., structured according to our normative model), moderate-structured,

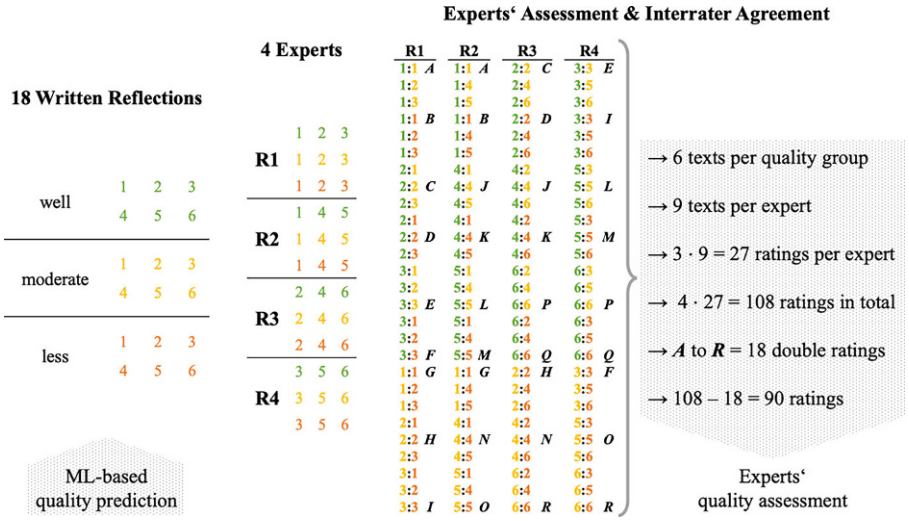

Fig. 3 Scheme of distribution of randomized selected written reflections to the experts, experts' ratings, double ratings for agreement among experts, and the experts' assessment for human-machine agreement. Each Arabic number (1 to 6) refers to a specific written reflection, the colors refers to the quality groups

and less-structured texts (quality groups). From each of these three groups, three randomly selected texts were given to four experts for observation and evaluation of classroom situations. In our case, physics seminar instructors in the second phase of teacher training programs were selected. These experts supervise novice teachers in their first years through seminars and frequent classroom observation and reflection meetings. Each expert had the task to compare each text of one quality group with all texts of the other two quality groups (see Fig. 3). The expert rating was guided methodically by a manual that was attached to the nine texts. This manual specified the texts to be compared. ("Compare text A with text B and decide which text is a more successful reflection"). No hierarchy of all texts has been made, but only the best of two texts was further considered. The experts were not told which aspects they should pay attention to during the comparison. They were asked to make decisions based on their experience in contact with novice teachers (intuitive-heuristic judgement). The 9 texts resulted in 3×6 (well-structured with moderate- and less-structured texts) $+ 3 \times 3$ (moderate-structured with less-structured texts) $= 27$ comparisons for each expert. The psychometric literature suggests that statistical (actuarial) judgements and decisions are on average more accurate compared to clinical judgements by experts (Grove et al. 2000). Hence, aggregating experts' judgements on the written reflections was considered a valuable strategy to approximate quality. The LOS was not shown to the experts, so that they could make their independent judgement. In total the four experts conducted $4 \times 27 = 108$ comparisons.

In order to be able to check the validity of this evaluation, two experts always received an identical text from each quality group. Thus, 18 comparisons were carried out twice (see Fig. 3, *A*–*R*). The interrater agreement was calculated using Cohens' kappa. Consequently, a total of 18 randomly selected texts (six per quality

group) were presented to the experts. If there is sufficient agreement among the experts, the agreement between the LOS based grouping into quality groups and experts' assessment is calculated for all of the 108 – 18 = 90 comparisons using Cohens' kappa once more. Basically, this is a true-false proof, whether experts' judgement which written reflection might be of higher quality than the other and ML based calculation which written reflection has a higher LOS are equal.

In order to evaluate the feasibility of the comparisons, the experts evaluate how difficult a comparison was on a five-point Likert scale after each comparison (from 1 = "very easy comparation" to 5 = "very difficult comparation", arithmetic scale mean = 2.5). Using this evaluation, we obtain an additional information about the qualitative comparability of written reflections.

5.5 Quality assessment using LOS based on the output of the ML model

To answer RQ 3 the reflection elements and the validated LOS of a written reflection are related to each other. The actual distribution represents the sentence wise analyzed reflection elements in a written reflection. The reasoning structure and the proportions of descriptive reflection elements (descriptions) and discursive elements (evaluations, alternatives, and consequences) can be extracted out of this. Correlation calculations of the ML model output and its modelling (relative and absolute proportions of the reflection elements, Level of structure) are compared to common influential constructs of text quality (e.g., essay length). Cohens' kappa was used as the measure of correlation and interpreted according to Landis and Koch (1977).

6 Sample, validation and findings

6.1 Sample characteristics

Of the $N = 110$ participants $N = 49$ were bachelor's students and $N = 43$ were master's students (pre-service physics teachers) of three German universities of three federal states. $N = 15$ further participants were in-service teachers from different locations in two federal states. 29 participants were female 78 were male. This is a rather typical distribution. Participants were aged between 19 and 49 years ($M = 25.3$, $SD = 5.27$). The mean text length of the written reflections was 743 words ($SD \approx 305$) with a range from 231 up to 1755 words.

6.2 RQ 1: Characteristics of written reflections

To discuss the representativeness of the written reflections in our sample, we first examine the percentage of reflection elements (Fig. 4) in order to estimate how frequently each reflection element is represented on average in the written reflections and compare it with our normative distribution. Most of the written reflections predominantly use the reflection element descriptions ($mean = 54.12\%$, $SD = 15.69\%$). The proportions of evaluations ($mean = 27.21\%$, $SD = 12.00\%$), followed by the formulated proportions of alternatives ($mean = 12.18\%$, $SD = 6.38\%$), are significantly

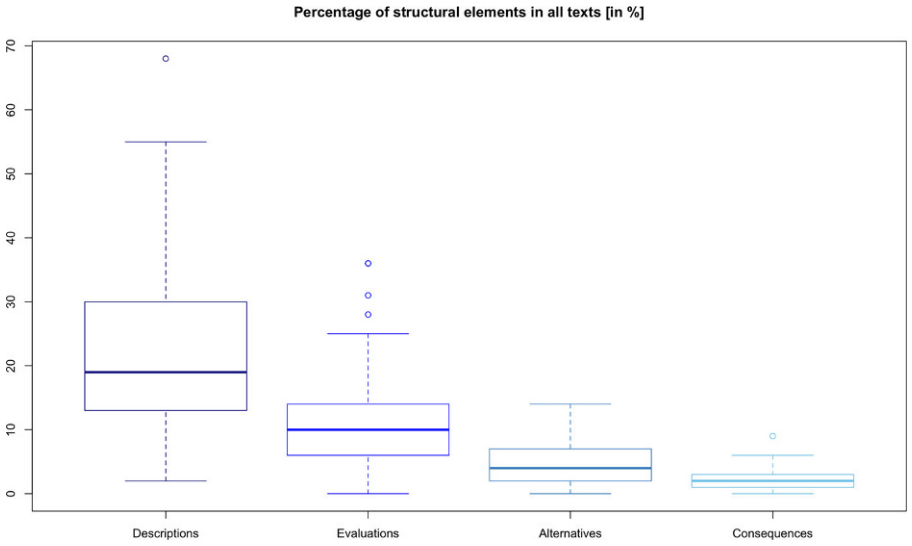

Fig. 4 Percentage of reflection elements in all Texts (in %)

lower ($p < 0.001$). Consequences comprised the smallest share of the reflection elements ($mean = 6.49\%$, $SD = 4.97\%$). Mean differences to our normative distribution (35% descriptions, 35% evaluations, 15% alternatives, 15% consequences) are +19.12% (descriptions), –7.79% (evaluations), –2.82% (alternatives), and –8.51% (consequences). All differences are also significantly. Thus, the teachers agree most with the proportion of alternatives formulated with our collaborative written reflection. Overall, the written reflections are descriptive, which is in line with findings of other researchers (Lai and Calandra 2007; Mena-Marcos and Tillema 2006).

6.3 RQ 2: Experts' ratings and ML based estimation

The calculated level of structure (LOS) of the $N = 110$ reflection texts is in the average 0.289 ($SD \approx 0.161$), values ranging from –0.064 to 0.703, and represents a normal distribution ($W = 0.991$, $p = 0.716$). The LOS of five texts is less than zero. The interval boundaries of the LOS for the less-structured texts are [–0.064 | 0.142], for the moderate-structured texts are [0.259 | 0.318], and for the texts of the well-structured quality group are [0.415 | 0.703] (see Fig. 5). Dividing our sample of $N = 110$ written reflections by quantiles each quality group is represented by $N = 22$ reflection texts. Significant differences in the lengths of the written reflections cannot be identified between the quantiles.

The four experts carried out $N = 18$ doubled ratings. The raters agreed in 12 cases and disagreed in 6 cases. From the 12 agreements 8 rated as predicted regarding the LOS and 4 as unexpected. It is of interest that the difficulty of the comparison is rated easier for the experts for the predicted agreements ($mean = 2.00$, $SD = 0.53$) than for the disagreements ($mean = 2.83$, $SD = 0.68$) or the unexpected agreements

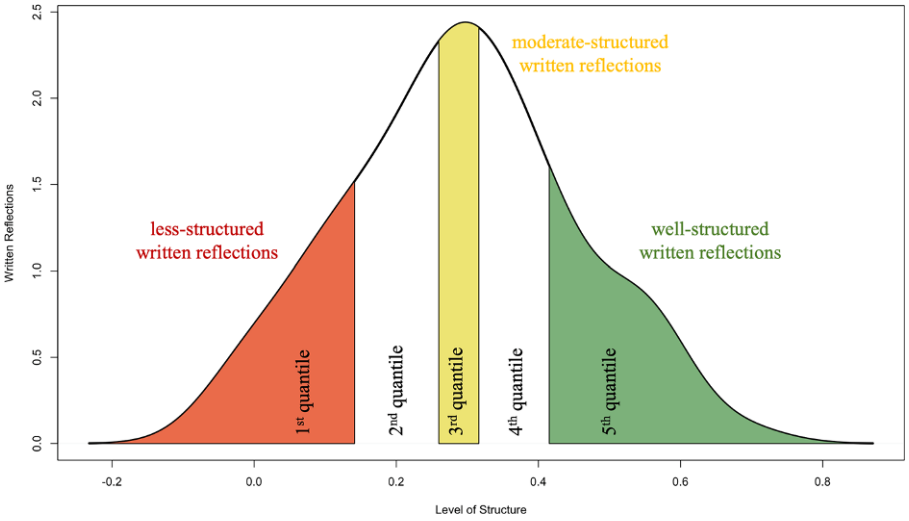

Fig. 5 Distribution of the LOS over all 110 written reflections and the three quality groups (*red*: less-structured, *yellow*: moderate-structured, *green*: well-structured)

(*mean* = 3.00, *SD* = 0.82). So, comparisons are rated most difficult when experts agree amongst each other but disagree with the ML model scores.

Regardless of the double ratings the overall agreement between experts' ratings and ML based estimation of $N = 108$ evaluations is 72.2% or $\kappa = 0.70^{***}$. Considering the doubled ratings (see Table 1), the human-computer-agreement ranks around 68.6% ($\kappa = 0.635^{***}$). ML based comparisons with reflection texts from the quality group of less-structured texts turn out to be more similar to the decision of human

Table 1 Experts' agreement with the ML based estimation & experts' evaluation of the difficulty of the comparisons

ML based estimation		Rater 1	Rater 2	Rater 3	Rater 4	Mean
Total	Agree (%)	85.2	66.7	63.0	59.3	68.6
	κ	0.823***	0.612***	0.573***	0.531***	0.635***
Less-mod	Agree (%)	88.9	77.8	77.8	88.9	83.4
	κ	0.842***	0.700***	0.700***	0.842***	0.772***
Mod-well	Agree (%)	77.8	44.4	44.4	11.1	44.4
	κ	0.700***	0.348**	0.348**	0.077	0.368*
Less-well	Agree (%)	88.9	77.8	66.7	77.8	77.8
	κ	0.842***	0.700***	0.571***	0.700***	0.703***
Difficulty of the comparisions	Total	2.92	2.04	3.22	1.85	2.51
	Less-mod	2.89	1.78	3.00	1.67	2.33
	Mod-well	3.63	2.33	3.89	2.00	2.96
	Less-well	2.33	2.00	2.78	1.89	2.25

*: $p < 0.05$
**: $p < 0.01$
***: $p < 0.001$

evaluators, while the comparison with texts from quality groups of moderate- and well-structured texts seems to be more difficult. These comparisons are also rated as more difficult by the experts. The accuracy of the human-computer-agreement is thus analogous to the experts' assessment of the difficulty of a comparison. In general, the experts for lesson observation and feedback seem to rate the written reflections of the quality groups according their quality similarly according to the LOS.

6.4 RQ 3: Quality indicators and modelled structure

In order to evaluate the influence of the text length of a written reflection in our sample, we first considered the relative proportions of the reflection elements because the absolute frequencies of the elements increase with text length anyway. We found that the text length in our sample was less related to the relative proportions of formulated descriptions, evaluations, alternatives, and consequences. However, a significant negative correlation was found between the percentage of descriptions and the other three reflection elements. Likewise, a significant positive correlation was found between the percentages of alternatives and consequences. In texts with a higher proportion of alternatives, more consequences could be identified and vice versa. Figure 6 illustrates these correlations as a correlation panel. The distributions of the variables (Text length & proportions of descriptions, evaluations, alternatives, and consequences) are shown as histograms on the diagonal. Above the diagonal, the Cohen's kappa coefficients are shown with the significance stars according to Landis and Koch (1977). Below the diagonal, the correlations are shown as scatter plots to give a more detailed impression.

Assuming that the correlation of length and quality also applies to written reflections, this analysis suggests that the quality of a written reflection cannot depend solely on the proportion of interpretative elements.

A second correlation matrix was calculated where the LOS was correlated with the absolute frequencies of the reflection elements to investigate whether quality correlates can be observed in this way. Because the LOS was derived amongst others from the percentages of the reflection elements, the LOS was analyzed in conjunction with the absolute frequencies of the formulated sentences which were classified as descriptions, evaluations, alternatives, and consequences. LOS and text length were found to be not significantly related to each other ($cor = 0.10$, $p = 0.28$). However, the identified positive correlation between alternatives and consequences also become clear in this representation ($cor = 0.30$, $p < 0.01$). In addition, there were positive correlations between the absolute frequencies of evaluations and the absolute frequencies of formulated alternatives and consequences, so all discursive reasoning elements. The impression from Fig. 6 with the significant negative correlations between descriptions and all other reflection elements is not found in the correlation panel in Fig. 7. Instead, the absolute frequencies of the reflection elements descriptions and evaluations correlate significantly positive. A possible influence of the level of structure is obvious in a significant negative correlation with the absolute frequencies of descriptions or the positive correlations with the frequencies of formulated alternatives.

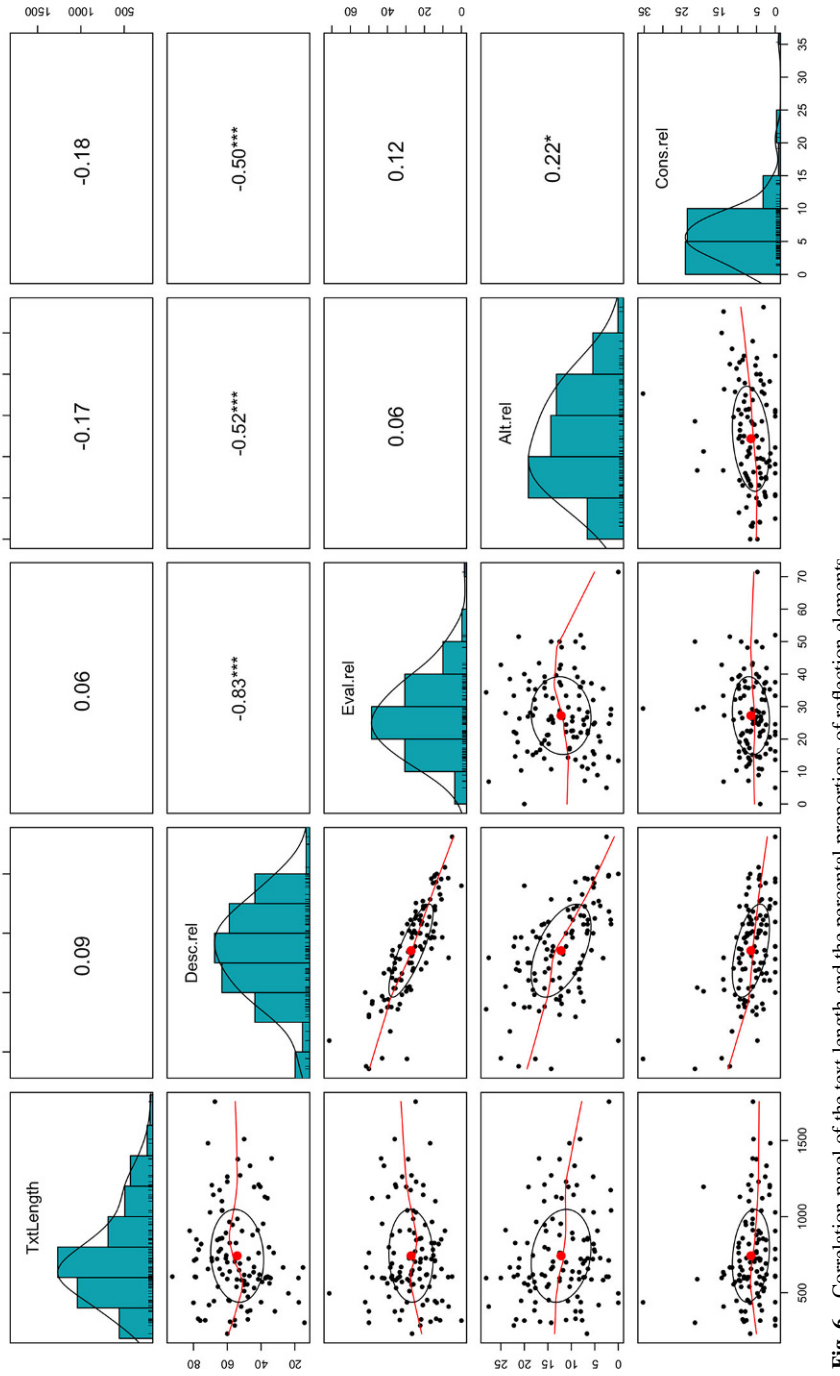

Fig. 6 Correlation panel of the text length and the percental proportions of reflection elements

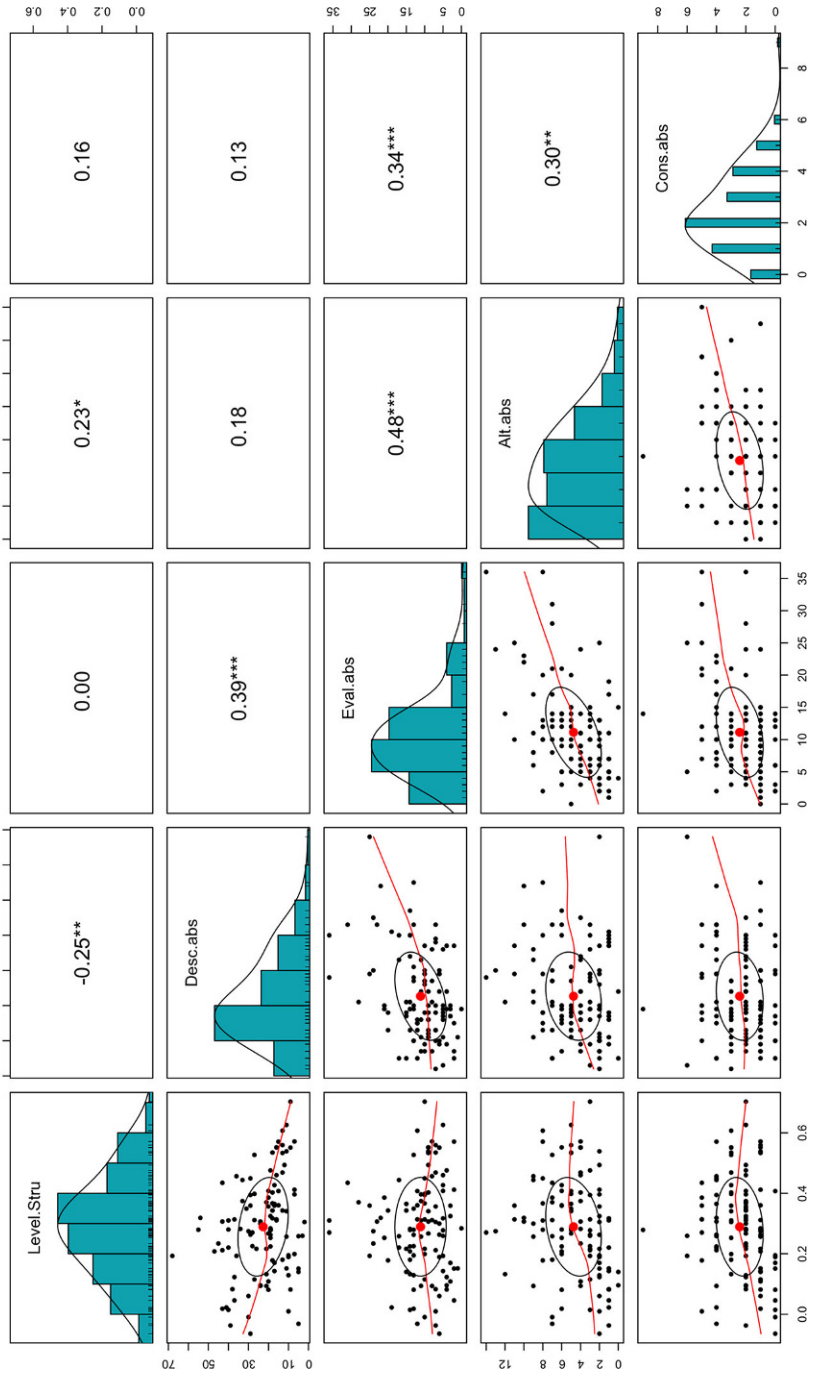

Fig. 7 Correlation panel of the LOS and the absolute number of formulated reflection elements

These data suggest that the level of structure (LOS) can be seen as an indicator of the quality of a written reflection. A higher LOS is related to a less descriptive text and written reflections with a higher proportion of formulated alternatives. Fewer evaluations seem to be associated with fewer descriptions, although evaluations may be necessary for the formulation of alternatives and consequences.

7 Summary

As of now, assessment of written reflections either relies on content analysis approaches or more data-centered approaches such as ML. In this study, we further examine potentials of ML-based approaches to the assessment of written reflection in physics education, in particular, to what extent the ML-based outputs such as textual classifications are related to experts' judgements on written reflections. Answering this question can facilitate researchers means to adopt ML models into intelligent tutoring applications and automated guidance, as is readily available today with ground-breaking applications such as ChatGPT. Three RQs guided our analyses:

7.1 RQ 1: Characteristics of written reflections

ML-based analysis of distribution of elements revealed that the written reflections in our sample are predominantly descriptive, as was found by other researchers in different contexts (Hatton and Smith 1995; Mena-Marcos et al. 2013) and are therefore valid. However, interindividual differences were notable. Participants formulated descriptions, evaluations, alternatives and consequences to varying degrees. It is remarkable but not unexpected that the proportion of descriptions is greater on average than the proportion of evaluations. Alternatives were also formulated less frequently, and the lowest proportion were consequences for one's own professional development. This finding implies the necessity to scaffold pre-service physics teachers' reflection-based reasoning processes. Computer-based feedback, that makes use of the here-employed models, might be a starting point. Guidance for pre-service physics teachers could single out the frequency of the elements, and suggest the inclusion of interpretative elements (see Kost 2019) if proportion is below average.

7.2 RQ 2: Experts' ratings and ML based estimation

To verify whether texts scaled higher in our ML-based metric are actually higher in quality according to expert assessments, four experts compared 18 reflection texts in a total of 108 comparisons and rated the difficulty of each comparison. For this (1) experts' agreement among each other ratings and (2) the agreement of experts' ratings with the ML based rating was calculated. Overall, the experts agreed in 72.2% ($\kappa = 0.70$***) with the ML based estimations, which can be seen as a substantial agreement (Landis and Koch 1977). If the double coding is considered, the number of agreements dropped slightly to 68.6% ($\kappa = 0.64$***). Experts declared the ratings with a less-structured text as relatively easy and ratings without less-

structured texts, meaning the comparisons of written reflections from the quality group of moderate- and well-structured texts as the most difficult. Whereas human-machine agreements for the quality groups moderate-structured and well-structured were lowest. However, human-machine agreements for the texts of the quality groups less-structured and well-structured were far better compared to the other groups. This result is expected, because these two extreme groups were the outermost of the five quantiles of the LOS. The comparison of texts in the quality groups of less-structured and moderate-structured texts reached the greatest human-machine agreement with 83.4% and a Cohens' kappa of 0.772***.

In summary, the human-human agreement as well as the human-machine agreement corresponds with the difficulty of comparisons estimated by the experts. It can be concluded that the level of structure (LOS) accords to experts' ratings and can consequently be regarded as a quality criterion or quality estimation of a written reflection (at least in extreme groups). The accuracy of the ML modelling seemingly makes a clear quality statement especially for less-structured reflection texts. Therefore, the ML based distinction between less-structured and moderate-structured reflection texts could be considered as part of a quality assessment of the written reflections, because they correspond to experts' ratings.

7.3 RQ 3: Quality indicators and modelled structure

Because the text length is known as a quality correlate in texts we focused on the relative proportions of the reflection elements. According to this we found significant negative correlations between the percentage of descriptive and discursive elements. For our written reflections it seems that the greater the proportion, especially of formulated descriptions, the fewer words of a text are spent on discursive elements (controlling for text length), such as evaluations, formulating alternatives and concluding personal consequences. Another significant correlation exists between the proportion of formulated alternatives and consequences. This means that in a text in which more alternatives are formulated percentage wise, more consequences are written (and vice versa). The proportion of formulated evaluations is not related to this proportionality. Furthermore, our data suggest that the text length has no notable influence on the relative distribution of reflection elements within a text. In case we consider a written reflection to be of higher quality, if it is less descriptive but contains more evaluations, alternatives and consequences, this means that longer texts do not automatically lead to better written reflections. This finding is contrary to other writing analytics findings where amongst others the text length represents text quality (Fleckenstein et al. 2020; Hatton and Smith 1995; Mena-Marcos et al. 2013).

Looking at the absolute frequencies of formulated descriptions, evaluations, alternatives and consequences in the text, the negative influence of the descriptions on the quality of the texts seems to decrease. In contrast, a significant positive correlation was found between absolute descriptions and evaluations. Further positive correlations were found for alternatives and consequences. Contrary to the comparison of the relative proportions, significant correlations between the discursive elements can be identified among the absolute frequencies. This attributes to the

fact that if a pre-service teacher includes more alternatives, she or he also includes more consequences overall. We conclude that a greater proportion of evaluations leads to more discursive reasoning, as this is accompanied with more alternatives and consequences being discussed. The positive correlations between the reflection elements and the absolute number of formulated descriptions could indicate that a certain minimum level of description of the observed situation is necessary for high-quality written reflection.

Calculating LOS presents a fast-and-frugal method for assessing the text quality. Focusing again on the calculated correlations between the LOS and the absolute frequencies of reflection elements, it becomes clear that a more precise structure that is consistent with the model does not only help to formulate significantly fewer descriptions, but consequently supports the use of discursive elements in written reflections. The modelling we propose with the LOS seems to represent indicators of the complex structure of written reflection quality and might be able to exclude the length of a written reflection at the same time.

8 Synopsis, limitations, and future perspectives

Our study shows that ML model scores can be used to automatically assess aspects of quality for written reflections. From a sentence-by-sentence assignment of the reflection texts into descriptions, evaluations, alternatives and consequences, both percentage shares and the position of the reflection elements in the text could be determined with the help of ML. From a collaborative normative reflection, a level of structure (LOS) could be assigned to each reflection text. This LOS was validated by four experts in observation and analysis of physics education.

From our ML based analysis of the written reflections, we could not identify connections between the text length and relative proportions of discursive reflection elements. Therefore, we argue that we found a measure within our LOS in which length as a typical quality indicator is extracted, so that other indicators in research on written reflections can come into focus. Measured by our proposed determination of a LOS, we can confirm that structured texts (according to the normative distribution that we proposed) also contain more discursive elements. The structure we have determined is directly related to the discursive elements of a written reflection. Nevertheless, a minimum level of formulated sentences seems to be relevant for a successfully structured and higher-quality reflective text. As a caveat we also posit that our normative distribution is one among many different possibilities. We cannot exclude the possibility that this particular distribution has blind spots and certain written reflections that are of high quality receive low scores because they do not attend to the normative distribution. Future research should examine different normative distributions and also the sequencing of reflection elements in the written reflections in greater detail. As such, our approach can be seen as a template for assessing text structure on a discourse element level (Stede and Schneider 2019). The reflection elements form as discourse entities that fulfill certain rhetorical and cognition-related function, e.g., the description accesses episodic memory and sets the ground for discursive processes that are important in reflection-related reasoning.

Our proposed analysis considers reflection quality and reflection structure according to Nowak et al. (2019) reflection-supporting model. Reviews of reflection-related writing show that most reflection-supporting models outline similar discourse elements (Poldner et al. 2014; Ullmann 2019). Hence, we suggest, our presented approach and assessment method can function as a template for these other models. Changing the discourse elements is merely a question of extending the coding process. Alternative reflection-supporting models also include reflection elements such as feelings or personal beliefs (Ullmann 2019). Research into these different reflection-supporting models would require some further training of the here-described ML model. It would be interesting to evaluate to what extent these different models yield similar findings regarding prominent reflection elements and correlations among elements. Our LOS as a quality correlate can be adopted in the novel contexts without difficulties.

Furthermore, we were able to confirm our method of LOS using (elementary) expert ranking. It is worth noting that the level of agreement between experts and ML based analysis is more difficult for well-structured reflection texts. Experts also rate these comparisons as more difficult. In this way, it must be limited that the structural analysis proposed by us has less significance especially for better structured reflection texts. Nevertheless, the informative value of our ML based structural analysis seems to be consistent with the quality of these texts, especially in the area of less-structured reflection texts. This provides further evidence that structure in reflective texts is an important basis for higher-quality reflective processes.

From a theoretical stance, it was widely recognized that multiple definitions for reflection coexist and it is difficult to discern based on mere philosophical/pedagogical arguments which perspective is most fruitful for supporting learning and professional development (Ullmann 2019). We suggest that ML-based approaches and our LOS in particular provide meaningful tools to advance reflective writing analytics based on empirical evidence. As such, ML and LOS can provide means to validate reflection-supporting models and even falsify aspects of them. For example, when certain reflection elements are not observed in practice, this would raise questions as for how useful these elements are. Data-driven discovery methods can play an important role to advance our understanding of reflection-related reasoning processes. With speech and language technologies, even interviews could be automatically transcribed and large-scale analysis of these corpora (big data) becomes readily possible. Empirically testable and quantifiable models might be developed on this basis.

Following the outlined definition of reflection of von Aufschnaiter et al. (2019), structured reflection-related reasoning might be a key component for effective professional development. For this reason, the proposed level of structure (LOS) according to the reflection-supporting model for reflection of Nowak et al. (2019) might be helpful to track pre-service teachers' development in reflection competence and to scaffolding/feedback written reflections. Lai and Calandra (2007) report: "both teacher educator and pre-service teacher participants strongly suggested the use of reflection-related online resources to provide in-time, on-demand mechanism to develop pre-service teachers' reflection-related knowledge base, and ultimately, to enhance their reflectivity development" (p. 78). Not only do pre-service teachers appreciate such scaffolding, but it also is effective for the development of reflec-

tion-related reasoning processes (Lai and Calandra 2007). The presented reflection-supporting model presents an important means to scaffold the process of writing reflection and establishing transparency of this task, which students demand (Lai and Calandra 2007).

We want to encourage other university instructors to feedback pre-service teachers' reflective practice regarding the reflection structure with the help of ML. Computer-generated scaffolds/feedback might contribute to this (e.g. Mientus et al. 2021). Especially for novice teachers who are not as successful in using the reflection-supporting model of Nowak et al. (2019) to structure their reflective process, the LOS might be meaningful for the quality of a written reflection. Therefore, we propose to use the ML-based modelling of the complex reflection process as a kind of filter to give automated feedback on the reflective structure to pre-service or novice teachers who are weak in reflective competences. Thus, university teachers have more resources for content-related feedback and can implement this in a more analytic and less holistic way.

Funding This research was funded by the German Federal Ministry of Education and Research and is part of the 'Qualitätsoffensive Lehrerbildung', grant number 01JA1816.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest L. Mientus, P. Wulff, A. Nowak and A. Borowski declare that they have no competing interests. The funders had no role in the design of the study; in the collection, analyses or interpretation of the study data; in the writing of the manuscript or in the decision to publish the results.

Ethical standards All procedures performed in studies involving human participants or on human tissue were in accordance with the ethical standards of the institutional and/or national research committee and with the 1975 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

Abels, S. (2011). *LehrerInnen als 'Reflective Practitioner': Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht*. Wiesbaden: VS.

Aeppli, J., & Lötscher, H. (2016). EDAMA – ein Rahmenmodell für Reflexion. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *34*(1), 78–97.

von Aufschnaiter, C., Fraij, A., & Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung. *HLZ*, *2*, 144–159. https://doi.org/10.4119/UNIBI/hlz-144.

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

Biernacki, R. (2014). Humanist interpretation versus coding text samples. *Qualitative Sociology*. https://doi.org/10.1007/s11133-014-9277-9.

Blomberg, G., Renkl, A., Sherin, M. G., Borko, H., & Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. *Journal for Educational Research Online*, *5*(1), 9.

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R., & McWilliams, M. (2017). Towards reflective writing analytics: rationale, methodology and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. https://doi.org/10.18608/jla.2017.41.5.

Cappell, J. (2013). *Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase* (Studien zum Physik- und Chemielernen, Vol. 146). Berlin: Logos.

Carlson, J., Daehler, K., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K., Cooper, R., & Friedrichsen, P. (2019). The refined consensus model of pedagogical content knowledge. In A. Hume, R. Cooper & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science*. Singapore: Springer. https://doi.org/10.1007/978-981-13-5898-2_2.

Chodorow, M., & Burstein, J. (2004). Beyond essay length. Evaluating e-rater's performance on Toefl essays. *ETS. Res. Rep.*, *2004*, i–38. https://doi.org/10.1002/j.2333-8504.2004.tb01931.x.

Christian, B. (2021). *The alignment problem. How can machines learn human values?* London: Atlantic Books. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6294690

Christof, E., Rosenberger, K., Köhler, J., & Wyss, C. (2018). *Mündliche, schriftliche und theatrale Wege der Praxisreflexion* (Beiträge zur Professionalisierung pädagogischen Handelns). Bern: Hep.

Darling-Hammond, L. (2012). *Powerful teacher education. Lessons from exemplary programs* (1st edn.). Jossey-Bass. http://gbv.eblib.com/patron/FullRecord.aspx?p=695739

van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teaching and Teacher Education*, *24*(2), 244–276. https://doi.org/10.1016/j.tate.2006.11.005.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.562462.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19.

Häcker, T. (2022). Reflexive Lehrer*innenbildung Versuch einer Lokalisierung in pragmatischer Absicht. In C. Reintjes & I. Kunze (Eds.), *Reflexion und Reflexivität in Unterricht, Schule und Lehrer:innenbildung* (pp. 94–116). Bad Heilbrunn: Klinkhardt.

Hatton, N., & Smith, D. (1995). Reflection in teacher education: towards definition and implementation. *Teaching and Teacher Education*, *11*, 33–49. https://doi.org/10.1016/0742-051X(94)00012-U.

Hey, T., Tansley, S., Tolle, K., & Gray, J. (2009). *The fourth paradigm: data-intensive scientific discovery*. Microsoft research.

Hume, A. (2009). Promoting higher levels of reflective writing in student journals. *Higher Education Research & Development*, *28*(3), 247–260.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics. Linguistic comprehension and production. In J. Hay, R. Bod & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–95). Cambridge: MIT Press.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). GPT-4 passes the bar exam. *SSRN*. https://doi.org/10.2139/ssrn.4389233.

Kember, D., Jones, A., Loke, A., McKay, J., Sinclair, K., & Tse (1999). Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. *International Journal of Lifelong Education*, *18*, 18–30. https://doi.org/10.1080/026013799293928.

Kori, K., Pedaste, M., Leijen, Ä., & Mäeots, M. (2014). Supporting reflection in technology-enhanced learning. *Educational Research Review*, *11*, 45–55. https://doi.org/10.1016/j.edurev.2013.11.003.

Korthagen, F. A., & Kessels, J. (1999). Linking theory and practice: changing the pedagogy of teacher education. *Educational Research*, *28*(4), 4–17.

Kost, D. (2019). *Reflexionsprozesse von Studierenden des Physiklehramts (Dissertation)*. Gießen: Justus-Liebig-University.

Krieg, M., & Kreis, A. (2014). Reflexion in Mentoringgesprächen – ein Mythos? *Zeitschrift für Hochschulentwicklung*, *9*(1), 103–117.

Kuckartz, U. (2022). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* (5th edn.). Weinheim: Beltz Juventa.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Lai, G., & Calandra, B. (2007). Using Online scaffolds to enhance preservice teachers' reflective journal writing. A qualitative analysis. *International Journal of Technology in Teaching and Learning*, *3*(3), 66–81.

Lai, G., & Calandra, B. (2010). Examining the effects of computer-based scaffolds on novice teachers' reflective journal writing. *Educational Technology Research and Development*, *58*(4), 421–437.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Leonhard, T. (2022). Reflexionsregime in Schule und Lehrerbildung – Zwischen guter Absicht und transintentionalen Folgen. In C. Reintjes & I. Kunze (Eds.), *Reflexion und Reflexivität in Unterricht, Schule und Lehrer:innenbildung* (pp. 77–93). Bad Heilbrunn: Klinkhardt. https://doi.org/10.35468/5969-05.

Leonhard, T., & Rihm, T. (2011). Erhöhung der Reflexionskompetenz durch Begleitveranstaltungen zum Schulpraktikum? Konzeption und Ergebnisse eines Pilotprojekts mit Lehramtsstudierenden. *Lehrerbildung auf dem Prüfstand*, *4*(2), 240–270.

Lindmeier, A. (2011). *Modeling and measuring knowledge and competencies of teachers. A threefold domain-specific structure model for mathematics*. Münster: Waxmann.

Mayring, P. (2022). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (13th edn.). Weinheim: Beltz.

Mena-Marcos, J., & Tillema, H. (2006). Studying studies on teacher reflection and action: An appraisal of research contributions. *Educational Research Review*, *1*(2), 112–132. https://doi.org/10.1016/j.edurev.2006.08.003.

Mena-Marcos, J., García-Rodríguez, M.-L., & Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *European Journal of Teacher Education*, *36*(2), 147–163. https://doi.org/10.1080/02619768.2012.713933.

Mientus, L., Klempin, C., & Nowak, A. (in press). *Reflexion in der Lehrkräftebildung – empirisch, phasenübergreifend, interdisziplinär*. Universitätsverlag Potsdam.

Mientus, L., Wulff, P., Nowak, A., & Borowski, A. (2021). ReFeed: Computerunterstütztes Feedback zu Reflexionstexten. In M. Kubsch, S. Sorge, J. Arnold & N. Graulich (Eds.), *Lehrkräftebildung neu gedacht. Ein Praxishandbuch für die Lehre in den Naturwissenschaften und deren Didaktiken* (p. 266). Münster: Waxmann. https://doi.org/10.25656/01:22414.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Nowak, A., Kempin, M., Kulgemeyer, C., & Borowski, A. (2019). Reflexion von Physikunterricht. In C. Maurer (Ed.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe* (pp. 838–841). Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Kiel 2018.

Onyx, J., & Small, J. (2001). Memory-work: the method. *Qualitative Inquiry*, *7*(6), 773–786.

Oser, F. K., & Heinzer, S. (2009). Die Entwicklung eines Qualitätskonstrukts zur advokatorischen Erfassung der Professionalität. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Eds.), *Lehrerprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (pp. 167–180). Weinheim: Beltz.

Oser, F., Heinzer, S., & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten. Chancen und Grenzen des advokatorischen Ansatzes. *Unterrichtswissenschaft*, *38*(1), 5–28.

Ovens, A., & Tinning, R. (2009). Reflection as situated practice: A memory-work study of lived experience in teacher education. *Teaching and Teacher Education*, *25*(8), 1125–1131. https://doi.org/10.1016/j.tate.2009.03.013.

Poldner, E., van der Schaaf, M., Simons, P. R.-J., van Tartwijk, J., & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teachacher Education*, *37*(3), 348–373.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, *50*(5), 1020–1049. https://doi.org/10.3102/0002831213477680.

Schön, D. A. (1983). *The Reflective Practitioner. How Professionals think in action*. Oxfordshire: Routledge.

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education*, *27*(2), 259–267. https://doi.org/10.1016/j.tate.2010.08.009.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2019). Automated essay scoring: writing assessment and instruction. In P. Peterson, E. Baker & B. McGaw (Eds.), *International encyclopedia of education* (3rd edn., pp. 20–26). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00233-5.

Sorge, S., Neumann, I., Neumann, K., Parchmann, I., & Schwanewedel, J. (2018). Was ist denn da passiert? *MNU Journal*, *6*, 420–426.

Sparks-Langer, G. M., Simmons, J. M., Pasch, M., Colton, A., & Starko, A. (1990). Reflective pedagogical thinking: how can we promote it and measure it? *Journal of Teaching Education*, *41*, 23–32. https://doi.org/10.1177/002248719004100504.

Stede, M., & Schneider, J. (2019). *Argumentation mining*. San Rafael: Morgan and Claypool.

Ullmann, T. D. (2017). Reflective writing analytics: empirically determined keywords of written reflection. In *LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM international conference proceeding series. (pp. 163–167).

Ullmann, T. D. (2019). Automated analysis of reflection in writing: validating machine learning approaches. *International Journal Artificial Intelligence in Education*, *29*(2), 217–257.

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., Stede, M., & Borowski, A. (2020). Computer-based classification of preservice physics teachers' written reflections. *Journal of Science Education and Technology*, *30*(1), 1–15. https://doi.org/10.1007/s10956-020-09865-1.

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2021). Stärkung praxisorientierter Hochschullehre durch computerbasierte Rückmeldung zu Reflexionstexten in der Physikdidaktik. *die hochschullehre*. https://doi.org/10.3278/HSL2111W.

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning—A case for pretrained language models-based clustering. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-022-09969-w.

Wulff, P., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2023). Enhancing writing analytics in science education research with machine learning and natural language processing—Formative assessment of science and non-science preservice teachers' written reflections. *Frontiers in Education*. https://doi.org/10.3389/feduc.2022.1061461.

Wyss, C. (2018). Mündliche, kollegiale Reflexion von videografiertem Unterricht. In E. Christof, K. Rosenberger, J. Köhler & C. Wyss (Eds.), *Mündliche, schriftliche und theatrale Wege der Praxisreflexion. Beiträge zur Professionalisierung pädagogischen Handelns* (pp. 15–49). Bern: Hep.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757.