Computer-Based Classification of Preservice Physics Teachers' Written Reflections

Peter Wulff¹ · David Buschhüter¹ · Andrea Westphal² · Anna Nowak¹ · Lisa Becker³ · Hugo Robalino³ · Manfred Stede³ · Andreas Borowski¹

Accepted: 8 September 2020 / Published online: 8 October 2020
© The Author(s) 2020

Abstract

Reflecting in written form on one's teaching enactments has been considered a facilitator for teachers' professional growth in university-based preservice teacher education. Writing a structured reflection can be facilitated through external feedback. However, researchers noted that feedback in preservice teacher education often relies on holistic, rather than more content-based, analytic feedback because educators oftentimes lack resources (e.g., time) to provide more analytic feedback. To overcome this impediment to feedback for written reflection, advances in computer technology can be of use. Hence, this study sought to utilize techniques of natural language processing and machine learning to train a computer-based classifier that classifies preservice physics teachers' written reflections on their teaching enactments in a German university teacher education program. To do so, a reflection model was adapted to physics education. It was then tested to what extent the computer-based classifier could accurately classify the elements of the reflection model in segments of preservice physics teachers' written reflections. Multinomial logistic regression using word count as a predictor was found to yield acceptable average human-computer agreement (F1-score on held-out test dataset of 0.56) so that it might fuel further development towards an automated feedback tool that supplements existing holistic feedback for written reflections with data-based, analytic feedback.

Keywords Reflection · Teacher professional development · Natural language processing · Machine learning

A source for preservice teachers' professional development during university-based teacher education are practical teaching experiences gained primarily during school placements (Zeichner 2010; Clarke and Hollingsworth 2002). Given the complex demands towards preservice teachers in school placements, external scaffolding was considered a major facilitator for professional growth (Grossman et al. 2009; Korthagen and Kessels 1999; Shulman and Shulman 2004). Such external scaffolding could comprise formative

assessment and feedback for preservice teachers (Korthagen and Kessels 1999; Scardamalia and Bereiter 1988; White 1994; Hattie and Timperley 2007). For example, researchers provided preservice teachers formative assessment for their written reflections on teaching enactments that eventually enabled them to think in a more structured way about their experiences to better analyze factors that influence their teaching (Poldner et al. 2014; Lai and Calandra 2010). However, it was recognized that currently formative assessment and feedback are mostly holistic, i.e., students receive feedback concerning the overall quality of their written reflection instead of more content-based, analytical feedback on the particular aspects of students' reflections that should be improved. Providing only holistic feedback may not be optimal for students' development to become better reflective practitioners (Poldner et al. 2014; Ullmann 2019). In addition, it does not scale well, given the amounts of students' teaching experiences in comparison with the available human resources that are required to create the feedback in the first place (Nehm et al. 2012, Ullmann 2017, 2019).

✉ Peter Wulff
peter.wulff@uni-potsdam.de

¹ Physics Education Research, University of Potsdam,
Karl-Liebknecht-Straße 24/25, 14476,
Potsdam-Golm, Germany

² Department of Educational Research, University of Potsdam,
Potsdam-Golm, Germany

³ Applied Computational Linguistics, University of Potsdam,
Potsdam-Golm, Germany

Recently, advances in computer technology enabled resource-efficient and scalable solutions to issues of formative assessment and feedback (Nehm and Härtig 2012; Ullmann 2019). Tools for formative assessment and feedback range from editing assistants that provide feedback on language quality (e.g., correct grammar, or coherence) to knowledge-oriented tools that identify the correctness of contents given an open-ended response format (Burstein 2009; Nehm and Härtig 2012). Computer-based and automated feedback tools that take textual data as input are also well-established tools in educational research (Shermis et al. 2019). These tools were implemented, among others, as automated essay scoring systems and can be based on methods of machine learning (ML) and natural language processing (NLP) that facilitate language modelling and automatization (Burstein et al. 2003; Shermis et al. 2019). For written reflections, such systems have been implemented in the medical profession or engineering (Ullmann 2019). In teacher education programs, such automated, computer-based feedback tools could be useful to extend formative feedback of preservice teachers' written reflections. These tools could facilitate teacher educators to supplement their feedback with content-based, analytical feedback and provide more opportunities for written reflection thus helping preservice teachers to think in a more structured way about their teaching experiences and better learn from them Ullmann (2017).

This study seeks to employ ML and NLP methods to develop a computer-based classifier for preservice teachers' written reflections that could in the long run identify elements of a reflection-supporting model in preservice physics teachers' written reflection. As such, we performed an experiment where these written reflections were disaggregated into segments of reflective writing according to a reflection model. A classifier was then trained to categorize the segments according to the elements of the reflection model. This computer-based classifier should function as a proof-of-concept, showing that it is possible to reliably classify segments in preservice teachers' written reflections that are based on a given reflection model.

Reflection in University-Based Teacher Education

Teachers' professional development is closely linked to learning from teaching-related experiences and analyzing these experiences with theoretical knowledge to develop applicable and contextualized teaching knowledge (Shulman 1986, 2001; Park and Oliver 2008). The development of applicable and contextualized knowledge was argued to be facilitated by reflecting one's own teaching enactments (Darling-Hammond

and Bransford 2005; Korthagen and Kessels 1999; Zeichner 2010; Carlson et al. 2019). Reflection can be defined as a "mental process of trying to structure or restructure an experience, a problem, or existing knowledge or insights." (Korthagen 2001, p. 58). As such, reflection is more encompassing compared with lesson analysis because it also addresses the professional development of the individual teacher and relates to their learning processes on the basis of experiences through teaching enactments (von Aufschnaiter et al. 2019).

However, reflecting one's teaching enactments is particularly difficult for preservice, novice teachers (Häcker 2019; Nguyen et al. 2014). Preservice teachers were found to reflect their teaching enactments rather superficially (Hume 2009; Davis 2006; Korthagen 2005). For example, it was often observed that novice teachers' written reflections were characterized by descriptive and evaluative writing without more advanced pedagogical reasoning involved or without reframing their experience through a different conceptual frame (Nguyen et al. 2014; Mena-Marcos et al. 2013; Kost 2019; Loughran 2002; Hatton and Smith 1995). Describing their teaching enactments in a neutral manner was particularly difficult for preservice teachers, where they oftentimes used value-laden and evaluative (mostly affirmative) words (Kost 2019; Poldner et al. 2014; Christof et al. 2018). The preservice elementary teachers in the study by Mena-Marcos et al. (2013) reflected mostly analytically imprecise on the level of habitual or descriptive reflection in their written reflections. Furthermore, most evaluations were affirming their own teaching, suggesting that mostly pre-existing beliefs were confirmed in these texts (see also, Christof et al. (2018)) and teachers remained in the personal realm (also, Clandinin and Connelly 2003). Also, preservice teachers tend to focus on describing and evaluating their behavior instead of using more advanced arguments and explicative reasoning, such as justification, causes, or transformation of personal professional development (Poldner et al. 2014; Mena-Marcos et al. 2013; Kost 2019).

Lin et al. (1999) proposed that reflective thinking could be facilitated, among others, through process models that help preservice teachers structure their thinking. Consequently, reflection-supporting models were developed that often adopt several phases that need to be involved in a well-structured reflection, such as observation, interpretation, reasoning, thinking about consequences, and enacting the consequences into practice (Korthagen and Kessels 1999; Mena-Marcos et al. 2013). Reflection-supporting models are most effective when they are grounded in the contents of students' reflections (Hume 2009; Loughran et al. 2001; Lai and Calandra 2010). A means to convey reflection-supporting models and improve preservice teachers' reflective thinking was by having preservice teachers write about

their practical teaching experiences and provide them feedback on aspects of the written reflection (Hume 2009; Burstein et al. 2003). Hume (2009) utilized written reflective journals in science teacher education. She reported that providing scaffolds for the preservice teachers, such as resources for teaching knowledge could improve the written reflections. However, without external scaffolding, she noticed, the preservice teachers were likely to fall back into lower levels of reflective writing. Bain et al. (1999) and others found that feedback on the level, structure, and content of preservice teachers' written reflection was the most effective strategy to promote their reflective writing (Bain et al. 2002; Zhang et al. 2019). Furthermore, Poldner et al. (2014) stressed that it was important for feedback to encourage preservice teachers to focus on elements of justification, dialogue, and transformative learning in their written reflections to become reflective practitioners.

The problem with recent approaches for scaffolding reflective writing in university-based teacher education is that they largely rely on holistic feedback (Poldner et al. 2014). Holistic feedback focuses on the level of students' written reflection. From a pragmatic perspective, holistic feedback is quick to create for teacher educators and thus time-effective. From a research perspective, the prevalence of holistic feedback can be linked to a lack of empirical, data-driven research on reflection where the focus remained on conceptual issues rather than how reflection can be taught and assessed (van Beveren et al. 2018). It has been suggested that holistic feedback is less supportive for preservice teachers because feedback should address particular contents and dimensions of the students' texts to enable the students to identify aspects of their reflections that they should strengthen in the future (Poldner et al. 2014; Ullmann 2019). To advance research on assessment of written reflections and formative feedback in teacher education, more analytical approaches are necessary. Analytical approaches are grounded in content analysis and would allow for a more structured and systematic feedback. For novice teachers in particular, analytical feedback can be more appropriate for their professional growth because the degree of concreteness particularly helps novices better understand their thinking process and model expert thinking processes which often remain tacit (Korthagen and Kessels 1999; Lin et al. 1999).

ML and NLP Approaches to Advance Analytical Feedback for Written Reflections

Analytical approaches to assessment of written reflections have been facilitated by advances in computer technology (Ullmann 2017). In particular, ML and NLP approaches enabled computer systems to learn from human texts

and perform classification tasks, such as detecting reflective paragraphs in students' texts (Ullmann 2019). ML approaches can be differentiated into supervised and unsupervised learning (LeCun et al. 2015). Supervised learning typically take human labelled data as input to fit weights to a mathematical model such that an output is modelled accurately (LeCun et al. 2015). Unsupervised learning rather represents structures and relationships that are present in the data without reference to human labelled data. For the purpose of classifying text segments, supervised learning is the method of choice.

NLP is particularly concerned with representing, processing, and even understanding and generating natural language through computer technology. NLP methods are widely employed in fields such as linguistics, where they can be used to automatically parse sentences into syntactic structure (Jurafsky and Martin 2014). Education research also embraced NLP methods to develop writing assistants, assessment systems, and intelligent computer-based tutors that often utilize methods of natural language understanding, which is a subset of NLP (Burstein 2009).

ML and NLP methods in combination are promising to develop classifiers and assessment tools for written reflections as well (Lai and Calandra 2010; Ullmann 2019). In fact, Ullmann (2019) trained an NLP-based classifier for segments in written reflections by health, business, and engineering students. It was found that five out of eight reflection categories could be detected with almost perfect reliability, and that accuracies of automated analysis were satisfactory, though, on average, 10% lower compared to human interrater assessment. For (science) teacher education programs, such computer-based assessment tools for preservice teachers' written reflections could serve as a means to raise quality and breadth of feedback and lower requirements in human resources.

The present research has the goal to develop a computer-based classifier for preservice physics teachers' written reflections on the basis of ML and NLP methods. This computer-based classifier should function as a proof-of-concept for a more involved feedback tool that would be capable of providing formative feedback on the structure of preservice physics teachers' written reflections so that the teachers can improve their ability to write about their teaching-related experiences, which is considered a first step in becoming a reflective practitioner (Häcker 2019). This feedback tool could supplement holistic feedback in a cost-efficient and less time-consuming manner (Ullmann 2019). The following research question guided the analyses: To what extent can a classifier be trained to reliably classify segments in preservice physics teachers' written reflections according to the reflection elements of a reflection-supporting model? Building a reliable computer-based classifier required us to adapt a reflection model to the present

purposes, manually label the written reflections according to the reflection elements of the model and train, and fit a classifier that achieves an acceptable agreement with human raters. In line with the analysis of teacher educators who found that teacher expertise is highly domain-specific and embedded in concrete situations (Berliner 2001; Shulman 1987; Clark 1995), the context of university-based preservice teacher education in physics was chosen. Physics is also considered a difficult and knowledge-heavy subject so that reflection-supporting models might particularly pay off (Sorge et al. 2018; Hume 2009).

Method

Adapting a Reflection Model for Written Reflections

In analyzing their scaffolding system for reflective journals, Lai and Calandra (2010) found that novice preservice teachers appreciated the given structure for their written reflection in the form of a reflection-supporting sequence model. Similarly, in this study, a reflection-supporting model was developed to help structure preservice physics teachers' written reflections. Korthagen and Kessels (1999) proposed a reflection-supporting model in the context of university-based teacher education that outlines the process of reflection in five sequential stages in a spiral model, where the last step of a former cycle comprises the first step of a subsequent cycle. According to this model, teachers start with recollecting their set objectives and recall their actual experiences. Afterwards, they outline essential aspects of the experience, reconstruct the meaning of the experience for their professional development, and identify larger influences that might have been relevant for the teaching enactment. Then, teachers generate alternatives for their actions with advantages and disadvantages for these alternatives. Finally, they derive consequences for their future teaching and eventually enact these action plans.

This rather broad model was adapted to the specific context of writing a reflection after a teaching enactment. This model was chosen as the theoretical framework because it is grounded in experiential learning theory and outlines essential steps to abstract from personal experience for the use of one's own professional development. The adapted model was particularly tailored to writing. In this line, a written reflection entails functional zones (as introduced in text linguistics by Swales (1990)) that each contributes to the purpose of the text. These functional zones pertain to elements of reflection that were gleaned from the model by Korthagen and Kessels (1999). The functional zones related to:

1. Circumstances of the taught lesson (Circumstances¹). Here, the preservice teachers were instructed to provide details on the setting (time, place), class composition (e.g., amount of students), and their set learning goals.
2. Description of a teaching situation in the lesson that the teachers wanted to reflect on (Description). Here, teachers were prompted to write about their actions and the students' actions.
3. Evaluation on how the teachers liked or disliked the teaching situation, and why (Evaluation). Here, teachers were held to judge their own performance and the students' performance.
4. Alternatives of action for their chosen activities (Alternatives). In this step, teachers should consider actions that they could have done differently to improve the outcomes.
5. Consequences they draw for their own further professional development (Consequences). In the final step, preservice teachers were encouraged to relate their reflection of the situation to their future actions.

We submit that these functional zones provide a scaffold for the process of written reflection. From argumentation theory, it is expected that valid conclusions (i.e., consequences for personal professional development in teacher education) rely on evidence (Toulmin 2003), which in the case of teaching situations should be students' responses to teaching actions (Clarke and Hollingsworth 2002). In the written reflections, preservice teachers are supposed to generate evidence through recall of essential environmental features (Circumstances) and the objective description of the teaching situation. The objective description of the teaching situation is a difficult task in and of itself given that perception is theory-laden so that structuring and external feedback are indispensable (Korthagen and Kessels 1999). It is furthermore important for teachers to abstract from the personal realm and integrate theory into their reflection (Mena-Marcos et al. 2013). This should be facilitated through evaluation of the teaching situation, generation of alternative modes of action, and outlining personal consequences.

From an empirical stance, the functional zones are constitutive of reflective writing as well. In particular, Hatton and Smith (1995) found that most of the labelled segments in their analyzed written reflections were descriptive and thus corresponded with the description phase in our model—eventually rendering Description (and also Circumstances)

¹The capitalized words Circumstances, Description, Evaluation, Alternatives, and Consequences refer to the elements of our reflection model in the remainder.

the most prevalent element in preservice teachers' written reflections (Poldner et al. 2014, also). Furthermore, Mena-Marcos et al. (2013) found that preservice teachers often included appraisals into their written reflections. This suggests that evaluation is another prevalent element in the reflections. Expert teachers with regard to deliberate reflection particularly used the descriptive phase to establish the context for exploring alternative explanations, which corresponds with the Alternative step in our model (Hatton and Smith 1995). Consequences are a particularly expert-like element in preservice teachers' written reflections, given that (genuine) thinking about consequences requires an openness for changing one's behavior in the future. Thinking about consequences should be trained because the goal of reflection is personal professional development (von Aufschnaiter et al. 2019; Poldner et al. 2014). Also, Ullmann (2019) found evidence that "future intention" (a category that essentially encapsulates Consequences) is well present in written reflections (though not assessed for the population of teachers) and could be accurately classified. With regard to distinctiveness of the five elements, it has been shown that elements like the ones in our model could be well distinguished (Ullmann 2019). What has been observed though was that Description and Evaluation are difficult to keep apart for novice, preservice teachers (Kost 2019).

Furthermore, the functional zones could be related to teachers' professional knowledge. Hence, the teachers' knowledge bases of content knowledge (CK), pedagogical knowledge (PK), and pedagogical content knowledge (PCK) were added to the model as another dimension (Nowak et al. 2018). Finally, the written reflection could be differentiated by the extent to which reasoning, e.g., justification, for an element of the reflection was present or not (see Fig. 1). However, the professional knowledge and the amount of reasoning present in the reflections were not a concern of this study, but rather only the contents of the functional zones (Circumstances, Description, Evaluation, Alternatives, Consequences, and Consequences), which are called elements of reflection in the remainder.

Fig. 1 Model for written reflection (Nowak et al. 2019)

Alternatives, and Consequences), which are called elements of reflection in the remainder.

The goal of the computer-based classifier was to accurately label segments as one of the elements of reflection. The elements of reflection seemed to be a reasonable choice for a classifier because these elements were not too specific, which would typically cause human-computer agreement to decrease (Nehm et al. 2012). Our prediction is that the classifier will be able to label segments in a written reflection according to these five elements.

Collecting Written Reflections in Preservice Teacher Education

Preservice physics teachers at a mid-sized German university were introduced to the reflection model during their final 15-week-long school placement. We decided to build the classifier upon preservice teachers reflections compared with expert teachers written reflections because the classifier was meant to be applied in teacher training programs. It was anticipated that the degree of concreteness of writing by preservice teachers would be different from expert teachers so that a classifier that was built upon expert teachers' written reflections might fail to perform well on preservice teachers' written reflections.

The preservice physics teachers were introduced to the reflection-supporting model and the five reflection elements in a 1.5-h seminar session. In this seminar session, preservice teachers learned the reflection elements and wrote a sample reflection based on an observed video recording of an authentic lesson. They were instructed to write their later reflections according to this reflection model. The contents of their written reflections were the teaching enactments that the preservice teachers made. Each preservice teacher handed in approximately four written reflections throughout the school placement. Writing about teaching enactments is a standard method in assessment of preservice teachers' reflections because it allows for careful thinking, rethinking, and reworking of ideas, and yields a product externalized from the person that is open for observation, discussion, and further development (Poldner et al. 2014). A writing assignment can be considered an open-ended, constructed response format. Constructed response formats, such as teachers' writing, were argued to be a viable method to assess reflective thinking (Poldner et al. 2014). Compared with selected-response formats like multiple-choice questions, writing can engage preservice teachers in a more ecologically valid and natural way of thinking and reasoning (Mislevy et al. 2002; Poldner et al. 2014).

Written reflections were collected from $N = 17$ preservice physics teachers (see Fig. 2). All preservice teachers were in their final school placement and were

Fig. 2 Procedure of data collection and human labelling of the data

recruited from two subsequent, independent semesters. In the course of this school placement, teachers were given multiple opportunities to hand in written reflections on their teaching enactments and receive expert feedback. Overall, $N = 81$ written reflections were collected.

To assess the predictive performance of the computer-based classifier, a held-out test data set was gathered in the same type of school placement 2 years later, until after building the computer-based classifier was finished. Classification of the held-out test data set and calculating agreement with human raters would serve as a means to evaluate the problem of overfitting of the model to the training dataset, given that multiple classifiers and feature configurations were considered. Instructions for the reflection-supporting model and the five elements of reflection were the same in this cohort compared with the other cohort. This cohort comprised $N = 12$ preservice physics teachers who were not related to the former cohort. $N = 12$ written reflections on the teaching enactments of the preservice teachers were collected.

In both cohorts, preservice teachers handed in their written reflections after going through a process of writing and editing these texts. The preservice teachers received a brief feedback by their mentors on breadth and depth for some of their written reflections.

Next, all the written reflections were labelled according to the elements of the reflection model (see Fig. 2). When teachers provided information, such as dates or meaningless information such that this was their second written reflection, these segments were labelled “irrelevant.” They were removed from further analyses because only a fraction (approximately 4%) of the segments was affected. Human interrater agreement for labelling the segments was used to assess to what extent teachers’ written reflections adhered to the reflection model and how reliably the elements could be identified by the human raters. Two independent human raters each labelled one written reflection of $N = 8$ different teachers (approximately 10% of the train and validation data sets, similar to Poldner et al. (2014)) twice. Each human rater read the entire written reflection. While reading, they

identified and labelled segments of the text according to the reflection elements of the model. A segment of the text was defined as a text fragment that pertained to one of the reflection elements and addressed a similar topic, such as making a case for the functioning of the group work during class. The human raters reached an agreement, as assessed through Cohen’s κ , of $\kappa = 0.74$, which can be considered substantial (Krippendorff 2019) and comparable with similar human coding tasks in research on written reflections (Poldner et al. 2014). We also considered a sentence-level coding with a subsample of the dataset, where agreement remained substantial with a noticeable decrease in agreement, $\kappa = 0.64$.

To evaluate where disagreements between the human raters occurred most often, a confusion matrix was created (see Table 1). Disagreements occurred most often with regard to Description, Circumstances, and Evaluation. Eleven percent of segments were misspecified by the two human raters with regard to these three elements. Alternatives and Consequences were never confused with Circumstances or Description, but occasionally with Evaluation.

After interrater agreement was assured, one human rater went through all written reflections and identified and labelled segments. This resulted in the following amount of segments per label: Circumstances, 759; Description, 476; Evaluation, 392; Alternatives, 192; Consequences, 147. The mean (SD) coding unit length was 2.0 (1.8) sentences, ranging from 1 to 29 sentences, with the longest being an extended description of a teaching situation. The mean (SD) number of words in a segment was 27 (29).

Building a Computer-Based Classifier for Written Reflections

Data Transformation and Feature Selection

To build the computer-based classifier, the written reflections have to be subdivided into segments which are transformed into features (predictor variables) that will allow the classifier to predict probabilities for each of the reflection

Table 1 Confusion matrix for human interrater agreement

Rater 1/rater 2	Circ.	Desc.	Eva.	Alt.	Conseq.	Σ
Circ.	133	15	0	0	0	148 (42%)
Desc.	1	87	3	0	0	91 (26%)
Eva.	1	12	36	0	6	55 (15%)
Alt.	0	0	1	32	1	34 (10%)
Conseq.	0	0	3	1	24	28 (8%)
Σ	135 (38%)	114 (32%)	43 (12%)	33 (9%)	31 (9%)	

elements for this segment. The choice of features is critical for classification accuracy and generalizability to new data (Jurafsky and Martin 2014). In this study, 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). Consequently, word count was used as a feature to represent a segment. Word count accounts for the words that are present in a segment through occurrence-based encoding in a segment-term matrix that is used to classify the segment (Blake 2011). In the segment-term matrix, rows represent segments and columns represent the words in the vocabulary. The cells are filled by the number of occurrences for a certain word in a given segment. This model ignores word order in the segments (bag-of-words assumption). On a conceptual level, word count combines multiple types of analyses, such as lexicon/topic (which terms are used?) or syntax/discourse (which terms co-occur in the same segment?) (Chodorow and Burstein 2004). Word count is often used as a starting point for model building, such that a classifier can be improved by comparing with other features, e.g., length of segment or lemmatized/stemmed word forms (Ullmann 2019).

To further assess performance sensitivity of the classifier to other feature configurations, further feature engineering was done in a later research question. In addition to the word count feature (see above), the following features were applied to all segments irrespective of category and were compared with each other:

1. **Meaningless feature (baseline model):** To get an idea to what extent a meaningless feature could accurately classify the segments, a baseline model was implemented with the feature of normalized (by segment length) vowel positions. Normalized vowel positions are not expected to relate to any category in a meaningful way so that the algorithm had no meaningful input for classification.
2. **Word count feature:** In this feature configuration (see above) the occurring words were encoded through a segment-term matrix. Punctuation and non-character symbols were removed from the documents.
3. **Some words feature:** In addition to the word count configuration from above, common words (called stop-words) were removed from the segments to reduce redundant and irrelevant information. Furthermore, words were included in a lemmatized form. Lemmatization transformed words into their base form, e.g., “was” will be mapped to “be.” However, modality was kept for verbs because this was considered an important feature for expressing alternatives or consequences, i.e., ideas in hypothetical form. Feature engineering of this kind lowers the dependency on idiosyncratic features of certain segments; however, it also removes potentially useful information, such as tense.
4. **Doc2Vec feature:** Finally, an advanced algorithm called Doc2Vec (Mikolov et al. 2013; Rehurek and Sojka 2010; Le and Mikolov 2014) was utilized to represent segments. Doc2vec is an unsupervised machine-learning algorithm that utilizes neural networks. In this algorithm, two complex models (continuous bag-of-words [CBOW], and skip-gram) are combined to retrieve word-vectors from the weight matrix after training a simple neural network in the CBOW and skip-gram steps. The word-vectors encode the context in which the respective word appears (CBOW) and predict the context based on the word (skip-gram). Alongside words, a special label for each document (segment) is also fed to the neural network, so that document-vectors that represent segments can also be retrieved from the weight matrix. This vector was used in the present study as a feature of the segment. Two particularly important hyperparameters for the Doc2Vec feature are window size and embedding dimensionality. Window size refers either to the size of the context (before and after the word) that is used to represent a word in the CBOW step, or to the size of context that a word predicts in the skip-gram step. Embedding dimensionality refers to the dimensionality of the space to which words and segments are mapped (Le and Mikolov 2014). With the (exploratory) purpose of the present study of testing the applicability of this algorithm to represent segments from written reflections, a (standard) window size of

15 was used alongside an embedding dimensionality of 300, which is considered a common configuration of hyperparameters for this model (Jurafsky and Martin 2014). Note that other configurations likely result in different values for performance metrics.

Classifier Algorithm

In the present study, supervised ML techniques were utilized in order to build the classifier. In the build phase of the computer-based classifier, segments were transformed into feature vectors and multiple classifier algorithms were fit with the goal of finding the model with the most performant combination of features and classifier (Bengfort et al. 2018). To find this model, the sample of $N = 81$ written reflections was split into training and validation datasets. To assess predictive performance of the most performant model for unseen data, it was finally fit to the held-out test dataset that consisted of $N = 12$ written reflections that were collected after the model building phase.

Classifier performance metrics included precision, recall, and F1-score for each category. Precision is the amount of correctly labelled positive data (true positives) for a category, given the amount of positive labelled data by the machine (true positives + false positives), and recall is the proportion of the number of items correctly labelled positive (true positives) given the number of items which should have been labelled positive (true positives + false negatives) (Jurafsky and Martin 2014). Due to the importance of both precision and recall, the harmonic mean of precision and recall is calculated, which is called F1-score. The harmonic

mean is more conservative compared with the arithmetic mean, and gives stronger weight to the lower of both values, i.e., it is shifted towards it. When F1-score is reported for entire models, the macro average and weighted average F1-scores across categories were used. Macro average refers to the average for precision, recall, and F1-score over the categories. Weighted average additionally accounts for the support for each category, i.e., number of segments for each category.

As a means to choose a classifier algorithm (see Fig. 3), multiple established classifiers were trained (Ullmann 2019) and used to predict the labels in the validation data set with default hyperparameter configuration (Bengfort et al. 2018). This approach indicates if some algorithms were more applicable to classify the segments. The choice for one classifier for further analyses can be justified on the basis of these results. Four classifier algorithms were implemented: Decision Tree Classifier, multinomial logistic regression, Multinomial Naïve Bayes, and Stochastic Gradient Classifier. The Decision Tree Classifier creates a hierarchical model to classify data. The input features (words) are used in a rule-based manner to predict the labels of the segments (Breiman et al. 1984). Decision Tree Classifier has beneficial attributes, such as explainable decisions or flexibility with input data. However, it is prone to generate overcomplex trees with the problem of overfitting. Multinomial logistic regression is a versatile model to build a classifier where weights are trained based on optimization methods to maximize a loss function (Jurafsky and Martin 2014). A softmax function is used to calculate the probabilities of any of the five categories. Multinomial logistic regression

Fig. 3 Logic of fitting the models

is specifically useful because it balances redundancy in predictor variables and has convenient optimization properties (convex function, i.e., it has one global minimum) (Jurafsky and Martin 2014). The Multinomial Naïve Bayes classifier builds on Bayes' rule where the posterior probability for a category of a segment is calculated through the prior probability of the label across all segments and the likelihood that, given the category, the features (e.g., words) occur. Yet, it is often infeasible to compute the evidence for any combination of words in a given segment so that the simplifying bag-of-words (or naïve) assumption is made, namely that the classification is unaffected if a certain word occurs in 1st or any other position (Jurafsky and Martin 2014). Multinomial Naïve Bayes has the particular advantage to perform well for small datasets but has less optimal characteristics when features are correlated. Finally, Stochastic Gradient Classifier is a classifier for discriminative learning where an objective function is iteratively optimized. Stochastic Gradient Classifier optimizes the evaluation of the loss function for linear classifiers, e.g., logistic regression.

To fit the classifiers, the programming language Python 3 (Python Software Foundation), the scikit-learn library (Pedregosa et al. 2011), and gensim (Rehurek and Sojka 2010) were used. Furthermore, the libraries spaCy (version 2.2.3) and nltk (Bird et al. 2009) were used to pre-process segments and extract features.

Results

Fit Different Classification Algorithms

Our RQ was the following: To what extent can a computer-based classifier be trained that reliably classifies a held-out test dataset with respect to the elements of reflection? In order to evaluate this RQ, multiple classifiers were used in a first step to predict the validation data. In this approach, word count was used as the feature to represent

the segments. Later on, feature engineering was considered (see Fig. 3).

Table 2 displays the values for precision, recall, and F1-score for the four classifiers and averages of these values. The upper part of the table displays the measures for precision, recall, and F1 for each category separately. In the bottom part of the table, macro and weighted averages of the score across categories are displayed. The Decision Tree Classifier had the lowest macro and weighted average F1-scores. In particular, the Decision Tree Classifier performed particularly poorly for Description and Evaluation as assessed through F1-scores. Multinomial logistic regression, Multinomial Naïve Bayes, and SGD Classifier performed comparably with each other in terms of macro and weighted average precision, recall, and F1.

The weighted values for the performance metrics are almost always higher compared with the unweighted values across classifier algorithms because the categories with the most support could be classified with the best performance. Furthermore, the calculated macro average precision values for the four respective classifiers appeared to be higher than or at least equal to the respective calculated macro averages for recall. The classifiers were better to return relevant against irrelevant results compared to returning most of the relevant results.

In sum, we proceeded using multinomial logistic regression for further analyses given the comparably good performance (see Table 2) and the advantageous convergence properties of regularized multinomial logistic regression (Jurafsky and Martin 2014).

Further Feature Engineering

As of now, word count was used as the feature to represent the segments. However, further feature engineering might be of value because on the one hand the training data set consisted of 1573 segments. On the other hand, the vocabulary size was 5774 words and the average number of words for a segment was only 28. Hence, the segment-term

Table 2 Precision (p), recall (r), and F1-score (F1) for the classifiers applied to our data set

	DecTreeClass.			LogReg			MultiNB			SGDClass.			Support
	p	r	F1	p	r	F1	p	r	F1	p	r	F1	
Circ.	0.73	0.79	0.76	0.78	0.91	0.84	0.83	0.88	0.86	0.79	0.91	0.85	160
Desc.	0.53	0.47	0.50	0.73	0.64	0.68	0.73	0.74	0.74	0.79	0.61	0.69	100
Eval.	0.46	0.50	0.48	0.60	0.66	0.63	0.52	0.76	0.61	0.56	0.63	0.59	62
Alt.	0.60	0.62	0.61	0.83	0.62	0.71	1.00	0.41	0.58	0.77	0.77	0.77	39
Cons.	0.75	0.55	0.63	0.77	0.52	0.62	0.94	0.48	0.64	0.72	0.55	0.62	33
Macro	0.61	0.59	0.60	0.74	0.67	0.70	0.80	0.65	0.69	0.73	0.69	0.70	
Weighted	0.63	0.63	0.62	0.74	0.74	0.74	0.78	0.75	0.74	0.75	0.75	0.74	

Table 3 Performance metrics for different classification algorithms as assessed through weighted average over categories

	1. Baseline	2. Word count	3. Some words	4. Doc2vec
Accuracy	0.48	0.71	0.71	0.65
Precision	0.45	0.72	0.71	0.65
Recall	0.48	0.71	0.71	0.65
F1-score	0.46	0.71	0.71	0.65

matrix was a very sparse matrix with most cells in the matrix being zeros. To reduce the dimensionality of the segment-term matrix and eventually improve classifier performance, further engineering of the features that represent the segments is usually considered.

The multinomial logistic regression algorithm was used to evaluate further feature engineering (see Fig. 3). To find a feature specification that best approximated human labels, the models with different features were compared on the basis of weighted averages for the performance metrics. The four feature specifications were increasingly involved with requirements in background rules and algorithms. The baseline model used position of vowels as a dummy to essentially encode no relevant information with regard to the categories. The model with the word count feature was the same model as before, utilizing the aforementioned segment-term matrix. Note that the F1-scores did not exactly match because one hyperparameter was changed to accommodate for the more complex Doc2Vec feature. The third model, called some words (see Table 3), removed further information (common words; grammar, such as tense) from the segments through additional rules and algorithms. The final model used the most advanced algorithm in this paper (Doc2Vec) to represent segments in a lower dimensional vector space.

Results are displayed in Table 3. As expected, the baseline model classified the segments poorly with a weighted, averaged F1-score of 0.46. When the classifier was provided information in the form of word count, the measures improved for the F1-score to 0.71. When redundant information was removed (model: some words), the F1-score stayed essentially the same. The Doc2vec model had a lower F1-score compared with the word count models of 0.65.

Table 4 Highest weighted words in vocabulary to indicate phase in reflection model

	1	2	3
Circ.	kannte (knew)	Lernziel (learning objective)	Klasse (class)
Desc.	Beobachtung (observation)	Durchführung (execution)	den (the)
Eva.	Eine (a)	gut (well)	Erwartungsgemäß (as expected)
Alt.	wäre (would)	würde (would)	können (can)
Conseq.	Konsequenz (consequence)	anders (different)	darstellen (display)

Properties of the Classifier

As a means to partly explain the decisions of the classifier, the most informative words for each topic were extracted from the word count model with frequency-based encoding. This was done by choosing the words with the highest estimated weights, so they would eventually impact the classification strongest when present in a segment. The words are plausible for most elements (see Table 4). For example, all words in Circumstances (Circ.) can be considered meaningful for this category, given that preservice teachers were asked to report on the circumstances of their teaching enactment and give context information on their learning objectives and class size. For Description, the first two words capture certain actions in science lessons, e.g., students observed some phenomena. For Evaluation, all words appear to be representative for evaluative text that particularly appraises the own teaching enactment. For Alternatives, the words represent the subjunctive mode (similar to the English conditional) in German, which indicates that the speech is about hypothetical matters. For Consequences, the words are indicative of writing about future action plans and the employment of different “forms” of practice.

Fitting the Classifier to the Held-Out Test Dataset

To better adjust for the risk of overfitting classifiers to the data, it is recommended to use the classifier that was trained with the train and validation datasets to classify segments in a held-out test dataset that was not relevant for any of the previous decisions regarding features and classifier algorithms. The multinomial logistic regression model where all words were included to represent segments was thus fit to the held-out test dataset because this model was comparably performant compared to the other algorithms and feature specifications. The multinomial logistic regression was retrained with the entire training and validation data.

While performance metrics for some categories reached an acceptable level (see Table 5), the performance metrics for Alternatives and Consequences were quite low. The weighted averages of performance metrics were for precision 0.58, recall 0.56, and F1 0.56. It is notable that

Table 5 Performance metrics for five-way multinomial logistic regression classifier

	LogRegression		F1
	<i>p</i>	<i>r</i>	
Circ.	0.58	0.52	0.55
Desc.	0.59	0.67	0.63
Eva.	0.54	0.50	0.52
Alt.	0.38	0.29	0.33
Conseq.	0.29	0.53	0.38

the highest count in the confusion matrix (compared over rows and columns) was always in the diagonal. This indicates that no fundamental problems resulted for labelling the element, e.g., where one element was always confused with another element. Furthermore, Circumstances, Description, and Evaluation tended to be more often confused with each other (40%). For Alternatives and Consequences, no such clear patterns could be observed. Alternatives and Consequences were confused often with Evaluation, but also with the other categories. Overall, the distributions of labels as assessed through row/column sums and percentages were comparable. In particular, Alternatives and Consequences were chosen lowest, and Circumstances, Description, and Evaluation were chosen most frequently (Table 6).

Discussion

Providing analytical feedback for preservice physics teachers' written reflections that are based on their teaching enactments can be a supportive element in teacher education programs to facilitate structured reflective thinking and promote preservice teachers' professional growth (Clarke and Hollingsworth 2002; Poldner et al. 2014; Lai and Calandra 2010; Lin et al. 1999). Ullmann (2017) wrote that "[r]eflective writing is such an important educational practice that limitations posed by the bottleneck of available

teaching time should be challenged." (p. 167). To widen this bottleneck, ML and NLP methods can help make formative feedback more scalable and extend research on teachers' reflections in an empirical, data-driven manner (van Beveren et al. 2018). Based on these methods, this study tried to explore possibilities of classifying preservice physics teachers' written reflections according to the reflection elements of a reflection model. A computer-based classifier was trained and evaluated with regard to performance in classifying segments of the preservice physics teachers' written reflections. Human interrater agreement was utilized to assess if the reflection elements could be reliably identified by human raters. Two independent human raters reached substantial agreement in classifying the segments, so that we considered training a computer-based classifier worth the effort (Nehm et al. 2012; Ullmann 2019). With regard to our RQ (To what extent can a computer-based classifier be trained to reliably classify teachers' written reflections based on predefined elements of a reflection model?), a two-step approach was pursued to train a classifier. First, different classification algorithms were trained on segments from the training dataset to predict labels for segments in the validation dataset. All classifiers were found to have acceptable F1-scores, particularly if compared with similar classification applications (Burstein et al. 2003; Ullmann 2017, 2019). Regularized multinomial logistic regression was considered the most suitable classifier for the present purposes (i.e., comparably high F1-score, stable optimization characteristics also for small samples, appropriate treatment of redundant features). Second, further feature engineering such as lemmatization or Doc2Vec was done and showed no particular advantage over the word count model that included all words in a given segment for classification. We suspect that the size of our dataset was insufficient for Doc2Vec to be applied well. Algorithms like Doc2Vec are generally expected to improve performance; however, they are trained on much larger datasets (Jurafsky and Martin 2014; Mikolov et al. 2013). Overall, the classifier based on word counts performed with an F1-score of 0.71 on the validation dataset, and with an F1-score of 0.56 on the held-out test data. Given the complexity of the problem (five-way classification) and the comparably small

Table 6 Confusion matrix for predicting elements in held-out test dataset. Rows, human rater; columns, computer-based classifier

Computer/human	Circ.	Desc.	Eva.	Alt.	Conseq.	Σ
Circ.	95	28	29	3	0	155 (22%)
Desc.	60	145	62	3	2	272 (39%)
Eva.	32	28	115	22	6	203 (29%)
Alt.	4	4	5	20	4	37 (5%)
Conseq.	2	2	8	5	17	34 (5%)
Σ	193 (28%)	207 (30%)	219 (31%)	53 (8%)	29 (4%)	

size of the datasets, the trained classifier can—to a certain extent and with much room for improvement—classify the elements of reflection in physics teachers' written reflections on the basis of ML and NLP methods.

A particular problem occurred with regard to human–human and human–computer interrater agreement about Circumstances, Description, and Evaluation. In fact, 11% of human–human comparisons and 40% of computer–human comparisons in Circumstances, Description, and Evaluation were confused. Thus, the confusion in the human interraters seemed to have spoiled the learning algorithm that imitated that behavior to an even larger extent. The reason for the confusion in these elements might stem from imprecision in students' written reflections. Kost (2019) gathered evidence in the context of analyzing preservice physics teachers' written reflections that students had difficulties to separate descriptive and evaluative texts in reflections. It was also reported that preservice teachers' written reflections are oftentimes analytically imprecise (Mena-Marcos et al. 2013). However, fuzziness with regard to the reflection elements and imprecise writing are certainly barriers to accurate computer-based classification. Categorizing text on a coarser grained level (e.g., differentiating only two or three reflection elements) could be a strategy to improve accuracy especially when classifying novice teachers' written reflections. As indicated in the “Method” section, it could be advisable to train the classifier on expert teachers' written reflections. Expert teachers might be more precise in differentiating the elements of the reflection-supporting model. This would enable the classifier to identify imprecise writing in the preservice physics teachers' written reflections.

Furthermore, the limited size of the train, validation, and test data restricted the performance of the classifier. This became particularly apparent in Alternatives and Consequences. The support (i.e., number of segments) for Alternatives and Consequences was likely insufficient for the model to accurately identify segments of these elements. Elements with more support systematically performed better compared to low-support categories. It is unlikely that Alternatives and Consequences could not be classified in general, as evidenced in the study of Ullmann (2019), who was able to identify intention (a category that is similar to Consequences) accurately. Furthermore, the held-out test dataset was only gathered until after training of the classifier was completed. Consequently, the held-out test dataset did not comprise a random sample of the entire dataset (which would be desirable). Even though the instruction for the reflection model was similar in the two cohorts that comprised the training/validation and test data, the variance within students is generally smaller compared with the variance between students. Drawing a random sample from the entire dataset might improve performance on test data.

Future Directions

In this study, some of the implementational details for designing a computer-based, automated feedback tool for preservice physics teachers' written reflection could be outlined. Further optimization of the classification performance is expected to result from extending the available training data, advancing features for representing the segments, and improving the classification algorithms. Extending the training data could be achieved through further data collection in the teaching placement. Advancing features could be achieved, among others, through trimming the vocabulary or encoding semantic concepts that could be used for classification (Kintsch and Mangalath 2011). Even techniques of automated exploration of optimal features (representation learning) have been developed that could serve a great means in the context of advancing understanding of written reflections (e.g., what features best capture the texts) (LeCun et al. 2015). Further improvements in classification algorithms are expected to come from methods of ensemble classification, where multiple classifiers can be combined into a stronger classifier. This would enable us to decompose the segments by certain aspects, such as tense, lexicon, latent semantic concepts, or verb mode, so that classification is more tailored to the linguistic details of the segments. As we have seen, some of the linguistic details manifest in the analyzed words that were important for classification. Another direction for improving classification algorithms are expected to come from deep learning approaches. For example, pre-trained language models have been successfully applied to classification of book abstracts (Ostendorff et al. 2019). Employing pre-trained language models to physics and science education could help share resources among universities and compensate for limited sample sizes in these fields.

Regarding the development of an automated feedback tool, these performance improvements would be critical. What remained unclear in this study with regard to developing an automated feedback tool is to what extent the preservice teachers actually would acknowledge feedback that is solely based on the structure of their written reflection. Some limitations have to be considered for such feedback. For example, it was not validated to what extent preservice teachers in this study might have omitted contextual information in their written reflections due to varying levels of expertise. Especially tacit and implicit knowledge is a well-recognized problem for preservice and also expert teachers (Hatton and Smith 1995; Berliner 2001). Knowledge and information that is not articulated poses a problem for structural feedback on prevalence of the reflection elements because an expert might be told that she/he did not describe the situation in sufficient length,

when in fact the essential aspects were mentioned concisely (Hatton and Smith 1995). In particular, for expert teachers, structural feedback therefore needs to be supplemented with content-based feedback on reflective depth. Structural feedback alone might be suitable in the beginning for preservice teachers to learn reflecting on the basis of the reflection-supporting model that was proposed in this study.

Overall, improving the performance of the presented computer-based classifier could advance empirically oriented and data-driven approaches to reflection research and reflective university-based teacher education (Buckingham Shum et al. 2017). Improvements in classification performance would enable educational researchers to design formative feedback tools that provide preservice teachers with information on their process of reflection. Such feedback could relate to the ordering of the elements of reflection in their texts or to the extent that they include higher order reasoning, such as Alternatives and Consequences. The pedagogical goal for the computer-based classifier and feedback tools is to help preservice teachers structure their thinking about their teaching experiences that allows them to reason about their teaching enactments in a focussed and systematic manner which should help them become reflective practitioners.

Funding Open Access funding enabled and organized by Projekt DEAL. This project is part of the “Qualitätsoffensive Lehrerbildung,” a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research. The authors are responsible for the content of this publication.

Data Availability Please send requests to corresponding author.

Code Availability Please send requests to corresponding author.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they do not have a conflict of interest.

Consent/ethical statement Informed consent was assured with subjects.

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

- Bain, J.D., Ballantyne, R., Packer, J., Mills, C. (1999). Using journal writing to enhance student teachers' reflectivity during field experience placements. *Teachers and Teaching*, 5(1), 51–73.
- Bain, J.D., Mills, C., Ballantyne, R., Packer, J. (2002). Developing reflection on practice through journal writing: impacts of variations in the focus and level of feedback. *Teachers and Teaching*, 8(2), 171–196.
- Bengfort, B., Bilbro, R., Ojeda, T. (2018). Applied text analysis with Python. O'Reilly.
- Berliner, D.C. (2001). Learning about and learning from expert teachers. *Int J Educ Res*, 35, 463–482.
- Bird, S., Klein, E., Loper, E. (2009). *Natural language processing with Python*. Beijing and Cambridge: O'Reilly.
- Blake, C. (2011). Test mining. *Annual review of information science and technology*, 45(1), 121–155.
- Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A. (1984). *Classification and regression trees*, 1st edn. Boca Raton: CRC Press.
- 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.
- Burstein, J. (2009). Opportunities for natural language processing research in education. In Gebulkh, A. (Ed.) *Springer lecture notes in computer science* (pp. 6–27). New York: Springer.
- Burstein, J., Marcu, D., Knight, K. (2003). Finding the WRITE stuff: automatic identification of discourse structure in student essays - Intelligent Systems. IEEE [see also IEEE Expert]. IEEE Intelligent Systems, pp. 32–39.
- Carlson, J., Daehler, K., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., Park, S., Rollnick, M., Sickel, A., Suh, J., Schneider, R., van Driel, J.H., Wilson, C.D. (2019). The refined consensus model of pedagogical content knowledge. In Hume, A., Cooper, R., Borowski, A. (Eds.) *Repositioning pedagogical content knowledge in teachers' professional knowledge*. Singapore: Springer.
- Chodorow, M., & Burstein, J. (2004). Beyond essay length: evaluating e-rater's performance on toefl essays.
- Christof, E., Köhler, J., Rosenberger, K., Wyss, C. (2018). *Mündliche, schriftliche und theatrale Wege der Praxisreflexion (E-Book): Beiträge zur Professionalisierung pädagogischen Handelns*, 1st edn. Bern: Hep Verlag.
- Clandinin, J., & Connelly, F.M. (2003). Personal experience methods. In Denzin, N.K., & Lincoln, Y.S. (Eds.) *Collecting and interpreting qualitative materials* (pp. 150–178). London: Sage.
- Clark, C.M. (1995). *Thoughtful teaching*. Teachers College Press: Columbia.
- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teach Teach Educ*, 18(8), 947–967.
- Darling-Hammond, L., & Bransford, J. (Eds.) (2005). *Preparing teachers for a changing world: what teachers should learn and be able to do*. New York: Wiley. 1. auflage edition.
- Davis, E.A. (2006). Characterizing productive reflection among preservice elementary teachers: seeing what matters. *Teach Teach Educ*, 22(3), 281–301.
- Grossman, P.L., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., Williamson, P.W. (2009). Teaching practice: a cross-professional perspective. *Teach Coll Rec*, 11(9), 2055–2100.

- Häcker, T. (2019). Reflexive Professionalisierung: Anmerkungen zu dem ambitionierten Anspruch, die Reflexionskompetenz angehender Lehrkräfte umfassend zu fördern. In Degeling, M., Franken, N., Freund, S., Greiten, S., Neuhaus, D., Schellenbach-Zell, J. (Eds.) *Herausforderung Kohärenz: Bildungswissenschaftliche und fachdidaktische Perspektiven* (pp. 81–96). Bad Heilbrunn: Verlag Julius Klinkhardt.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Rev Educ Res*, 77(1), 81–112.
- Hatton, N., & Smith, D. (1995). Reflection in teacher education: towards definition and implementation. *Teach Teach Educ*, 11(1), 33–49.
- Hume, A. (2009). Promoting higher levels of reflective writing in student journals. *Higher Education Research & Development*, 28(3), 247–260.
- Jurafsky, D., & Martin, J.H. (2014). Speech and language processing. Always learning. Pearson Education, Harlow, 2. ed., pearson new internat. ed. edition.
- Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in cognitive science*, 3(2), 346–370.
- Korthagen, F.A. (2001). *Linking practice and theory: the pedagogy of realistic teacher education*. Mahwah: Erlbaum.
- Korthagen, F.A. (2005). Levels in reflection: core reflection as a means to enhance professional growth. *Teachers and Teaching*, 11(1), 47–71.
- Korthagen, F.A., & Kessels, J. (1999). Linking theory and practice: changing the pedagogy of teacher education. *Educ Res*, 28(4), 4–17.
- Kost, D. (2019). Reflexionsprozesse von Studierenden des Physiklehramts: Dissertation at Justus-Liebig-University in Gießen.
- Krippendorff, K. (2019). Content analysis: an introduction to its methodology. Sage, Los Angeles and London and New Delhi and Singapore and Washington DC and Melbourne, fourth edition edition.
- Lai, G., & Calandra, B. (2010). Examining the effects of computer-based scaffolds on novice teachers' reflective journal writing. *Educ Technol Res Dev*, 58(4), 421–437.
- Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents: Proceedings of the 31 st International Conference on Machine Learning. 32.
- LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lin, X., Hmelo, C.E., Kinzer, C., Secules, T. (1999). Designing technology to support reflection. *Educ Technol Res Dev*, 47(3), 43–62.
- Loughran, J. (2002). Effective reflective practice: in search of meaning in learning about teaching. *J Teach Educ*, 53(1), 33–43.
- Loughran, J., Milroy, P., Berry, A., Gunstone, R., Mulhall, P. (2001). Documenting science teachers' pedagogical content knowledge through PaP-eRs. *Res Sci Educ*, 31, 289–307.
- Mena-Marcos, J., García-Rodríguez, M.-L., Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *Eur J Teach Educ*, 36(2), 147–163.
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781v3.
- Mislevy, R.J., Steinberg, L.S., Almond, R.G. (2002). Design and analysis in task-based language assessment. *Lang Test*, 19(4), 477–496.
- Nehm, R.H., Ha, M., Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *J Sci Educ Technol*, 21(1), 183–196.
- Nehm, R.H., & Härtig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *J Sci Educ Technol*, 21(1), 56–73.
- Nguyen, Q.D., Fernandez, N., Karsenti, T., Charlin, B. (2014). What is reflection? A conceptual analysis of major definitions and a proposal of a five-component model. *Medical Education*, 48(12), 1176–1189.
- Nowak, A., Ackermann, P., Borowski, A. (2018). Rahmenthema "Reflexion" im Praxissemester Physik [Reflection in pre-service physics teacher education]. In A. Borowski, A. Ehlert, H. Prechtel (Eds.) *PSI Potsdam* (pp. 217–230). Potsdam: Universitätsverlag Potsdam.
- Nowak, A., Kempin, M., Kulgemeyer, C., Borowski, A. (2019). Reflexion von Physikunterricht [Reflection of physics teaching]. In C. Maurer (Ed.) *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe [Science education as a basis for professional and social participation]*. Jahrestagung in Kiel 2018 (p. 838). Regensburg: Gesellschaft für Didaktik der Chemie und Physik.
- Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., Gipp, B. (2019). Enriching BERT with Knowledge Graph Embeddings for Document Classification. arXiv:1909.08402v1.
- Park, S., & Oliver, J.S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Res Sci Educ*, 38(3), 261–284.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *J Mach Learn Res*, 12, 2825–2830.
- 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. *Eur J Teach Educ*, 37(3), 348–373.
- Python Language Reference: version 2.7. Retrieved from <http://www.python.org>.
- Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (Eds.) *Proceedings of the LREC 2010* (pp. 45–50).
- Scardamalia, M., & Bereiter, C. (1988). Development of dialectical processes in composition. In Olson, D.R. (Ed.) *Literacy, language, and learning*. Cambridge: Cambridge Univ. Press.
- Shermis, M.D., Burstein, J., Higgins, D., Zechner, K. (2019). Automated essay scoring: writing assessment and instruction.
- Shulman, L.S. (1986). Those who understand: knowledge growth in teaching. *Educ Res*, 15(2), 4–14.
- Shulman, L.S. (1987). Knowledge and teaching: foundations of the new reform. *Harv Educ Rev*, 57(1), 1–23.
- Shulman, L.S. (2001). Appreciating good teaching: a conversation with Lee Shulman by Carol Tell. *Educ Leadersh*, 58(5), 6–11.
- Shulman, L.S., & Shulman, J.H. (2004). How and what teachers learn: a shifting perspective. *J Curric Stud*, 36(2), 257–271.
- Sorge, S., Neumann, I., Neumann, K., Parchmann, I., Schwanewedel, J. (2018). Was ist denn da passiert? *MNU Journal*, 6, 420–426.
- Swales, J.M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge Univ. Press.
- Toulmin, S. (2003). *The uses of argument*. Cambridge: Cambridge University Press. updated ed. edition.
- Ullmann, T.D. (2017). Reflective writing analytics: empirically determined keywords of written reflection: LAK '17 Proceedings of the Seventh International Learning Analytics & Knowledge Conference. In: ACM International Conference Proceeding Series, pp. 163–167.
- Ullmann, T.D. (2019). Automated analysis of reflection in writing: validating machine learning approaches. *Int J Artif Intell Educ*, 29(2), 217–257.

- van Beveren, L., Roets, G., Buysse, A., Rutten, K. (2018). We all reflect, but why? A systematic review of the purposes of reflection in higher education in social and behavioral sciences. *Educational Research Review*, 24, 1–9.
- von Aufschnaiter, C., Fraij, A., Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung: 144–159 Seiten / Herausforderung Lehrer_innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion, Bd. 2 Nr. 1 (2019): Herausforderung Lehrer_innenbildung - Ausgabe 2.
- White, E.M. (1994). Teaching and assessing writing. Jossey-Bass Publishers.
- Zeichner, K.M. (2010). Rethinking the connections between campus courses and field experiences in college- and university-based teacher education. *J Teach Educ*, 61(1-2), 89–99.
- Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsumura, L.C., Howe, E., Quintana, R. (2019). eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 9619–9625.

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