frontiers in Education

ORIGINAL RESEARCH
published: 25 August 2020
doi: 10.3389/feduc.2020.00149

Automated Scoring of Teachers' Pedagogical Content Knowledge – A Comparison Between Human and Machine Scoring

Andreas Wahlen[1]*, *Christiane Kuhn*[1], *Olga Zlatkin-Troitschanskaia*[1], *Christian Gold*[2], *Torsten Zesch*[2] and *Andrea Horbach*[2]

[1] Chair of Business and Economics Education, Faculty of Law and Economics, Johannes Gutenberg University Mainz, Mainz, Germany, [2] Language Technology Lab, Faculty of Engineering, University Duisburg Essen, Duisburg, Germany

OPEN ACCESS

Edited by:
Susan M. Brookhart,
Duquesne University, United States

Reviewed by:
Mark Shermis,
American University of Bahrain,
Bahrain
Libby Gerard,
University of California, Berkeley,
United States

***Correspondence:**
Andreas Wahlen
anwahlen@uni-mainz.de

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: 08 April 2020
Accepted: 28 July 2020
Published: 25 August 2020

Citation:
Wahlen A, Kuhn C,
Zlatkin-Troitschanskaia O, Gold C,
Zesch T and Horbach A (2020)
Automated Scoring of Teachers'
Pedagogical Content Knowledge –
A Comparison Between Human
and Machine Scoring.
Front. Educ. 5:149.
doi: 10.3389/feduc.2020.00149

To validly assess teachers' pedagogical content knowledge (PCK), performance-based tasks with open-response formats are required. Automated scoring is considered an appropriate approach to reduce the resource-intensity of human scoring and to achieve more consistent scoring results than human raters. The focus is on the comparability of human and automated scoring of PCK for economics teachers. The answers of (prospective) teachers (N = 852) to six open-response tasks from a standardized and validated test were scored by two trained human raters and the engine "Educational SCoRIng Toolkit" (ESCRITO). The average agreement between human and computer ratings, κ_w = 0.66, suggests a convergent validity of the scoring results. The results of the single-sector variance analysis show a significant influence of the answers for each homogeneous subgroup (students = 460, trainees = 230, in-service teachers = 162) on the automated scoring. Findings are discussed in terms of implications for the use of automated scoring in educational assessment and its potentials and limitations.

Keywords: automated scoring, natural language processing, convergent validity, constructed responses, economics, pedagogical content knowledge

INTRODUCTION

Teaching a subject requires teachers to make the structure and meaning of the learning content accessible to learners, taking into account their individual learning prerequisites and needs (Kersting et al., 2014; Wilson et al., 2018). Pedagogical content knowledge (PCK) is considered a crucial facet of teaching competence (Shulman, 1986). To validly assess PCK, performance-based tasks with open-response formats are required (Alonzo et al., 2012; Zlatkin-Troitschanskaia et al., 2019), where test takers can describe their instructional approaches to teaching situations (Shavelson, 2009; Liu et al., 2016). The scoring of open responses by human raters is a resource-intensive process (Dolan and Burling, 2012; Zhang, 2013) and can lead to inconsistencies in the test scores due to personal rater biases, which limits objective, reliable and valid measurement (Bejar, 2012; Liu et al., 2014).

Computer-based methods of natural language processing (NLP) are used to analyze and subsequently score language-related features of texts in digital form (Burstein et al., 2013). Automated scoring is subdivided into the scoring of essays by "automated essay scoring" (AES) and the scoring of short-response texts by "automated short answer scoring" (ASAS) (Riordan et al., 2017). Automated scoring is considered an approach to reduce the resource intensity of scoring and achieve more consistent scoring results (Shermis et al., 2013; Zhang, 2013; Almond, 2014; Burrows et al., 2015). Differences between human and computer-based scorings may exist due to personal and dataset-related influences, for instance, gender or response length, or because of limitations of computer-based modeling (Bridgeman et al., 2012; Ramineni et al., 2012a,b; Perelman, 2014; Zehner et al., 2018).

In the few studies on the automated scoring of assessments, a positive correlation between automated and human scoring as well as a moderate to substantial agreement between human raters and computers was found in the domains of mathematics and sciences (Kersting et al., 2014; Wilson et al., 2017).

Similar studies on automated scoring are not available for teachers' PCK in the domain of economics. In this paper, we address the question of how comparable the automated and human scoring of PCK for economics teachers is. In particular, the research focus lies on the invariance of machine scoring as a validity criterion.

In this paper, we present the current research on the comparison of automated and human scoring with particular consideration of the aspect of invariance. The automated scoring of PCK is based on the responses of (teacher education) students, trainees, and in-service teachers (N = 852) from six open-ended items of a standardized test with text vignettes (Kuhn et al., 2016). The written responses were independently scored by two trained raters. The automated scoring was performed using the "Educational SCoRIng Toolkit" (ESCRITO) framework, Zesch and Horbach (2018). Next, we present analyses of the interrater agreement between human and machine scoring as well as of a single-sector variance, showing the influence of the responses for each homogeneous subgroup (students = 460, trainees = 230, in-service teachers = 162) on the automated scoring. We conclude with a critical consideration of the use of computer-based scoring in educational assessment.

STATE OF RESEARCH IN AUTOMATED SCORING

Comparability of Human and Machine Scoring

The comparison of automated and human scorings provides important evidence about the validity of written responses (Attali, 2013). In Page (1966), the quality of N = 138 essays of high school students in grades 8–12 was scored by four human raters and one machine rater. The correlation matrix indicated a moderate to strong positive correlation between all raters ($0.44 \leq r \leq 0.61$), and the correlation between automated and human scoring was as high as the correlation between the two human raters.

In Ramineni et al. (2012a), 3,000 essays from the graduate record examinations (GRE) Test were analyzed using the "electronic essay rater" (e-rater) engine by ETS. The results show a strong positive correlation between automated and human scoring for both task types (113 issue and 139 argument tasks) and across three tested scoring models (r_{issue} = 0.78; $r_{argument}$ = 0.79). All three e-rater models showed a substantial agreement with the human scoring for both task types ($0.73 \leq \kappa_w \leq 0.77$).

Ramineni et al. (2012b) also used the e-rater to analyze about 4,000 essays ($N \geq$ 152,000 responses) written for the Test of English as a Foreign Language (TOEFL). Again, the findings for both task types (38 independent and 38 integrated tasks) showed a strong positive correlation ($r_{Independent}$ = 0.75; $r_{Integrated}$ = 0.73) and a substantial agreement (κ_w = 0.70) between the automated and human scorings.

Evanini et al. (2015) used the written responses of N = 3,385 secondary school students from the TOEFL Junior Test. The results also showed a strong positive correlation ($0.65 \leq r \leq 0.67$) across all test items as well as substantial agreement ($0.62 \leq \kappa_w \leq 0.65$) between automated and human scoring.

Liu et al. (2014) conducted a meta analysis focusing on the accuracy of ASAS engines, indicating a moderate to almost complete agreement between automated and human scoring (e.g., Basu et al., 2013; Gerard and Linn, 2016; Drolia et al., 2017; Lee et al., 2019). For example, Leacock and Chodorow (2003) used the ETS scoring engine c-rater to analyze responses to open-ended tasks in mathematics and reading comprehension. The average agreement between c-rater and human rater for both domains was substantial, with κ_w = 0.73.

In the Automated Student Assessment Prize (ASAP) trials, Shermis (2015) evaluated the ASAS of response texts (N = 25,683) based on middle school students' content knowledge in different domains. To compare the summative scoring performance of the ASAS to that of trained human raters, eight scoring engines were used for automated scoring. While the human raters on average reached an almost complete level of agreement (κ_w = 0.89) with values ranging from $0.75 \leq \kappa_w \leq 0.96$, the level of agreement between the scoring engines was lower. Here, an average substantial agreement was achieved (κ_w = 0.72), with values ranging from $0.60 \leq \kappa_w \leq 0.82$ (Shermis, 2015).

In contrast to previous findings from the AES evaluation (Shermis, 2014), the results of the ASAS evaluation showed that the scoring engines used did not achieve the same degree of agreement as the human raters. Shermis (2015) sees an explanation for this result in the complexity of correct response options, given that human writers can produce many variations of words or phrases (which can all constitute a similarly correct response) that are easily recognized by human raters but can be easily disregarded by scoring engines. This can lead to systematic biases in scoring, which affect fairness.

In Germany, Zehner et al. (2016) analyzed the accuracy of automated scoring of N = 41,990 short responses from the 2012 Program for International Student Assessment (PISA) in

Germany. For all ten items, the agreement between automated and human scoring ranged from $0.46 \leq \kappa \leq 0.96$. Horbach et al. (2018) used a German version of the ASAP dataset and found a substantial agreement between computers and humans ($\kappa_w = 0.67$).

Computer-Generated Scoring of Pedagogical Content Knowledge

There are only a few studies that specifically investigate the automated scoring of teachers' knowledge. In the domain of mathematics, Kersting et al. (2014) analyzed the potentials of automated scoring as an alternative to human scoring of teacher responses. They used the Classroom Video Analysis instrument to measure teachers' usable knowledge. Teachers (N = 238–249) were shown 13–14 video clips from real mathematics lessons in three subject areas. Across all three areas, the computer scoring of usable teaching knowledge in mathematics is strongly positively correlated ($0.77 \leq r \leq 0.91$) and moderately consistent ($0.51 \leq \kappa_w \leq 0.55$) with the scorings of two human raters, indicating that the two scoring methods measure similar constructs.

Wilson et al. (2017) focused on the comparison of computer and human scoring of teachers' PCK in science, technology, engineering and mathematics (STEM). Teachers were shown video vignettes and asked to analyze the shown classroom situations in written form, using six dimensions of PCK (Stuhlsatz et al., 2016). The analyses confirmed a substantial agreement between humans and computers ($\kappa = 0.77$). Studies of this kind on teachers' knowledge in other domains such as economics are currently not available.

Invariance of the Automated Scoring

There are variations in the scoring results between human and automated scoring of the same responses (Condon, 2013). The lack of invariance of statistical parameters in automated scoring poses a great challenge for the interpretation and use of test results and can be attributed to person-related and response-related factors (Bejar et al., 2016).

Based on N = 132,347 essays from the GRE and TOEFL tests, differences between the e-rater scoring and two human scorings were identified by Bridgeman et al. (2012) with regard to the personal factors language, gender and ethnicity. The comparison of the scoring results of 10 language groups shows that test takers from China and Korea consistently receive higher scores in the automated scoring ($d_{Chinese} = 0.21$; $d_{Korean} = 0.10$). In contrast, Arabic, Hindi, and Spanish speakers receive higher scores from the human raters ($d_{Arabic} = -0.19$; $d_{Hindi} = -0.18$; $d_{Spanish} = -0.11$), indicating that Asian languages might be preferred by the e-rater. However, to test this assumption, other Asian languages, for example, Japanese, need to be included in the study. Moreover, human-human correlations ($0.66 \leq r \leq 0.76$) in the gender and ethnic groups are consistently lower than human-computer correlations ($0.75 \leq r \leq 0.81$), indicating that when predicting a score for only one person, the e-rater score should be preferred to the score determined by a human rater. There is a tendency for African American men and women to receive higher scores from humans than from the e-rater ($d_{male} = -0.22$; $d_{female} = -0.18$). The differences in other groups such as participants of Asian American, Hispanic and Indian American descent are quite small ($2.67 \leq M_{human} \leq 4.12$; $2.52 \leq M_{e-rater} \leq 4.04$).

Liu et al. (2016) also investigated whether differences between automated scoring and human scoring of certain subgroups can be determined, for example, based on the factors gender and language. Eight items with 379–1,922 answers from the field of natural sciences were evaluated using the scoring engine c-rater-ML[1]. This scoring showed no significant differences for all items with regard to gender. Human rater gave higher scores to men in three items, and for women in one item, but the differences were not significant, and the effect sizes were small ($d = 0.26$), indicating no consistent differences between human and automated scoring (Liu et al., 2016). With regard to language, for both automated and human scoring, only one item showed a significantly higher value for students who speak English as their first language in contrast to those who learned it as a second language ($d = 0.36$). Overall, the differences between the scoring methods were quite small.

Based on N = 400 essays from middle school students in the United States, Wind et al. (2018) examined whether human rater effects in training sets have an impact on the quality of scores generated by AES engines. Using the scoring engine Intelligent Essay Assessor (IEA), they focused on three rater effects: severity, centrality, and inaccuracy. The results are based on four training conditions that reflect these effects (most common, second most common, second least common, and the combination of the most common and the no-effect condition), and they showed that the AES engine tends to reflect these training conditions in their scoring results: severity ($1.75 \leq \text{av}_{human} \leq 2.76$; $1.75 \leq \text{av}_{IEA} \leq 2.75$)[2], centrality ($0.67 \leq SD_{human} \leq 0.87$; $0.64 \leq SD_{IEA} \leq 0.78$), and inaccuracy ($0.39 \leq r \leq 0.76$).

Homogeneity as a Further Variance Factor

Horbach and Zesch (2019) have recently been discussing other variance factors that influence automated scoring. One of these factors is whether the training data was collected from a homogeneous group of learners. To enable scoring engines to score items, they are trained beforehand with a set of answers that were previously scored by human raters. Horbach and Zesch (2019) assume that if a scoring engine learns only from the answers of a homogeneous group (e.g., students, trainees or in-service teachers), this also affects the variance of the computer scoring. We therefore focus on two research questions (RQs):

(1) How comparable is the automated and human scoring of teachers' PCK in economics?

(2) How do homogeneous groups influence the automated scoring of teachers' PCK in economics?

[1] C-rater-ML is a scoring engine developed by ETS that is typically applied to short answers ranging from a few words to a short paragraph (Loukina and Cahill, 2016).

[2] Average rating (av).

METHOD

Pedagogical Content Knowledge Instrument

The data is based on the answers to a standardized and validated paper-pencil test by Kuhn (2014) for assessing teachers' PCK in economics. The long version of the test consists of nine open-ended and eight closed situation-based tasks (Kuhn et al., 2016); the short version of six open-ended and five closed tasks (Kuhn et al., 2020). The closed tasks in multiple-choice (MC) format require the cognitive processes of application and analysis, the open-ended tasks additionally require the process of creation (of new ideas, products) (Kuhn et al., 2016). The didactic demands are reflected in aspects of lesson planning and reacting to students' statements in three content areas: sales, buying processes and principles of economics (**Figure 1**). To ensure objective results, a coding manual was developed together with experts from teacher training practice (Kuhn, 2014, **Table 1**).

All responses were available in German and were analyzed accordingly. The language skills of the student teachers and (prospective) teachers were not assessed, as they are assumed to have similar levels of test-relevant language skills due to their similar educational backgrounds.

Sample

The assessment of teachers' PCK in its long form (17 items) was conducted in 2011 (Kuhn, 2014). The target group ($N = 338$) included university students of business and economics education, trainees and in-service teachers of economics. For discriminant validation, the contrast group ($N = 142$) consisted of students of economics (without a teaching background) as well as trainees and in-service teachers with subjects other than economics (**Table 2**).

A further assessment of teachers' PCK using the short version of the test (11 items) was carried out in 2018 as part of the ELMaWi[3] project. The total sample ($N = 372$) comprised students of business and economics education, trainees and in-service teachers of economics (Kuhn et al., 2020, **Table 2**).

The 2011 and 2018 samples were similar in terms of age structure. As the level of education increased, the average age increased as well. Compared to 2011, the percentage of female students in business and economics education was higher in 2018, whereas the percentage of female in-service teachers was lower. The average school leaving examination grade (Abitur), an indicator of general cognitive abilities, was similar for the students of 2011 (2.4) and those of 2018 (2.3)[4]. In addition, other personal factors such as general cognitive abilities, self-efficacy and ambiguity tolerance were assessed. The analyses showed that there were either no or only weak correlations between these measured factors and the participants' PCK (general cognitive abilities: 0.014; self-efficacy: −0.008; ambiguity tolerance: 0.213**; **$p < 0.01$) (Kuhn et al., 2020). Therefore, it is questionable whether these factors may also affect the automated scoring of PCK, and they are consequently not included in the following analyses.

In this paper, we refer to a sample of $N = 852$ participants and $N = 5,112$ responses from 2011 to 2018. Due to possible cohort effects that might affect the scoring, the samples from 2011 to 2018 will be scored and analyzed both separately and together.

Scoring and Analysis Procedure

The open-ended responses were transcribed and coded by two human raters (Kuhn, 2014; Kuhn et al., 2020). In 2011, 62% of the open-ended responses (randomly selected) were double coded, whereas in 2018, all of the open-ended responses were double coded.

The ESCRITO for short responses, developed by Zesch and Horbach (2018), was used for the automated scoring. ESCRITO is a publicly available general-purpose scoring framework based on DKPro TC (Daxenberger et al., 2014). The basic principle is that a supervised machine learning classifier is trained using features extracted from a training data set of labeled data and then, on this basis, scores new, unseen test data; i.e., it tries to learn the scoring behavior of a human annotator. We used the annotated labels of two experts as gold standard; in cases where they diverged, we always used the first experts annotation. The scoring engine thus analyses each response set and compares it to previously analyzed responses and their scores.

Educational SCoRIng Toolkit has been shown to achieve state-of-the-art performance in various settings (Zesch et al., 2015; Horbach et al., 2017; Riordan et al., 2017). For our experiments, a separate classification model was trained for each item using training data related to that item. We used a standard feature set of lexical features, i.e., we extracted the top 10,000 token n-grams (uni- to tri-grams) and character n-grams (bi- to four-grams). Lexical features cover important words and word groups indicative of a correct or incorrect response while character features target sub-word units to take into account orthographic and morphological variance. We trained a Support Vector Machine (SVM) classifier using the SMO algorithm provided by WEKA (Witten et al., 1999).

Due to the limited amount of data, we did not divide the data into training and test data but performed 10-fold cross-validation instead, where 90% of the data was used to train a model which was then tested on the remaining 10% of items. This was done repeatedly until each item was tested once.

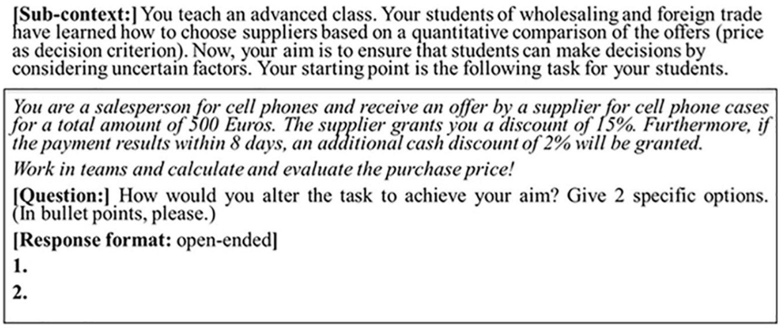
FINDINGS

Comparability of Human and Automated Scoring

With regard to the RQ1, the results show an almost complete agreement between the two human raters for all samples (2011: $\kappa_w = 0.87$; 2018: $\kappa_w = 0.91$; 2011/2018: $\kappa_w = 0.89$) (**Table 3**). For all items, a substantial to almost complete agreement is achieved ($0.65 \leq \kappa_w \leq 0.97$). The average scoring agreement between

[3] Assessing Subject-Specific Competences in Teacher Education in Mathematics and Economics, funded by the German Federal Ministry of Education and Research (BMBF) (https://www.eng.elmawi.de).

[4] In Germany, the grading system ranges from 1 to 6, with 1 being the best grade and 6 being the worst grade.



[Sub-context:] You teach an advanced class. Your students of wholesaling and foreign trade have learned how to choose suppliers based on a quantitative comparison of the offers (price as decision criterion). Now, your aim is to ensure that students can make decisions by considering uncertain factors. Your starting point is the following task for your students.

You are a salesperson for cell phones and receive an offer by a supplier for cell phone cases for a total amount of 500 Euros. The supplier grants you a discount of 15%. Furthermore, if the payment results within 8 days, an additional cash discount of 2% will be granted.

Work in teams and calculate and evaluate the purchase price!

[Question:] How would you alter the task to achieve your aim? Give 2 specific options. (In bullet points, please.)

[Response format: open-ended**]**

1.

2.

FIGURE 1 | Example item C15 "Comparison of offers" (lesson planning, content area: buying processes) (Kuhn, 2014).

TABLE 1 | Extract from the coding rules for item C15.

Code	Explanation *(examples)*
2	**Two of the following aspects:** Adding **further offers** from suppliers (currently no decision possible) ID 171: "I would specify different offers and the students would have to explain why they have chosen a certain supplier." Including **qualitative aspects** (e.g., long-term and trusting business relationship, adherence to delivery dates, environmentally friendly production) ID 196: "Suppliers who offer different levels of reliability should be included." **Fit with wholesale and foreign trade ("mobile phone seller")** ID 328 "reference to profession (wholesale and foreign trade)"
1	Only one aspect of the above-mentioned aspects
0	No, wrong or vague answers: "Take uncertain factors into account" [Here, an explanation/example is required]. "Work in pairs to calculate the purchase price with and without a cash discount"

TABLE 2 | Description of the 2011 and 2018 samples (Kuhn, 2014; Kuhn et al., 2020).

	2011			2018		
Target group *N* = 710	***N* = 338**			***N* = 372**		
	N	**Age (*SD*)**	**Female share**	***N***	**Age (*SD*)**	**Female share**
Students of business and economics education	176	24.3 (3.0)	55%	226	25.9 (4.2)	66%
Trainees of economics	109	31.0 (5.0)	57%	58	29.9 (5.6)	53%
In-service teachers of economics	53	42.3 (9.8)	55%	88	40.4 (9.4)	35%
Contrast group *N* = 142	***N***	**Age (*SD*)**	**Female share**			
Students of economics	58	23.4 (3.2)	44%			
Trainees and in-service teachers of subjects other than economics	84	38.0 (7.9)	48%			

the human raters and ESCRITO is lower, but still moderate to substantial ($0.59 \leq \kappa_w \leq 0.65$).

Figures 2, **3** show graphically to which extent the weighted kappa coefficient between human and human as well as human and computer varies between items and samples. For example, items A5, C13, and C15 achieve the highest agreement between human and automated scores in all three test runs (2011, 2018, 2011/2018) ($0.60 \leq \kappa_w \leq 0.77$). The values for A2, A4, and B10 were lower, but still moderate and increasing ($0.47 \leq \kappa_w \leq 0.63$)[5] in all three test runs (2011, 2018, and 2011/2018) (**Table 3**). This was illustrated by the fact that the degree of agreement increased with an increasing amount of (training) data (Heilman and Madnani, 2015). The results suggest that the automated and human scoring

[5] Landis and Koch (1977) have defined different value ranges for kappa with respect to the degree of agreement. Values greater than 0.60 can thus be used to represent a substantial to almost complete agreement beyond randomness. Values smaller than 0.40 represent a slight to weak match and values between 0.40 and 0.60 can be used to represent a fair to moderate agreement beyond randomness (Fleiss et al., 2003; Sim and Wright, 2005). According to Sim and Wright (2005), the definition of such reference values is arbitrary. McHugh (2012) has determined ranges of values which, according to the author, offer a more logical interpretation. However, these differ only slightly from Landis and Koch's definition and also adhere to the limit of 0.60, which is based on a sufficient agreement indication. Therefore, in this paper, the interpretation according to Landis and Koch is maintained.

TABLE 3 | Quadratically weighted kappa coefficients between human-human (H-H) and human-machine (H-M).

Item	2011 (κ_w)			2018 (κ_w)			2011 and 2018 (κ_w)		
	N^a	H-H	H-M	N^a	H-H	H-M	N^a	H-H	H-M
A2	465	0.94	0.62	375	0.92	0.49	840	0.93	0.53
A4	409	0.91	0.63	375	0.93	0.47	784	0.92	0.59
A5	450	0.79	0.65	375	0.94	0.75	825	0.86	0.69
B10	459	0.97	0.50	375	0.97	0.38	834	0.97	0.55
C13	420	0.65	0.77	375	0.95	0.76	795	0.80	0.75
C15	390	0.93	0.71	375	0.78	0.66	765	0.85	0.73
Item Average		0.87	0.65		0.91	0.59		0.89	0.64

[a] *N: unprocessed items are not included in the calculation.*

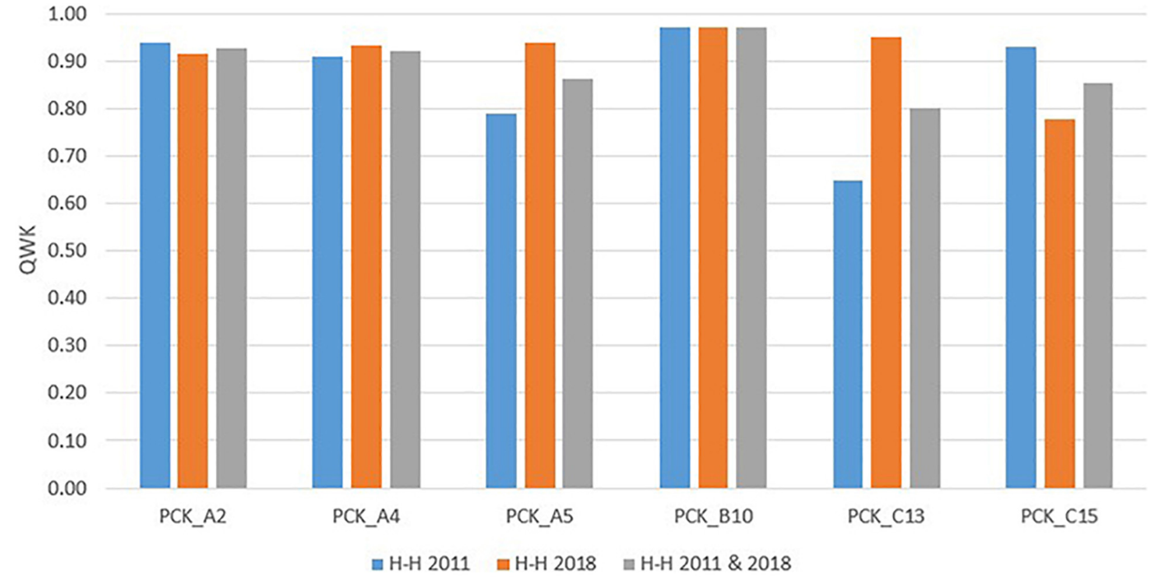


FIGURE 2 | Quadratically weighted kappa (QWK): Human-human.

of the PCK measure the same construct and differs between the items.

The PCK test also shows differences in the scoring agreement between the human and automated scoring results (**Figures 2**, **3**). **Figure 2** shows that the scoring agreement between human raters remains relatively constant across all datasets. Only item C13 shows a substantial difference between the 2011 and 2018 datasets. It can be assumed that this development is related to the subsequent revision and expansion of the coding manual that the human raters consult for their scoring process. **Figure 3** shows that the greatest differences between human-human agreement and human-machine agreement were found for items A2 and B10, while the agreement values for items A5, C13, and C15 were converging. With regard to the item features (**Table 4**), the partially low discriminatory power ($0.27 \leq r_{it} \leq 0.45$) and item difficulty (0.39–0.66) might be a possible cause for this result (Kelava and Moosbrugger, 2012; Kuhn, 2014; Horbach and Zesch, 2019).

With regard to the first RQ, the automated scoring of PCK is generally comparable to human scoring. It has the potential to deliver reliable scorings, which can have a positive effect on the validity of the data quality.

Invariance of the Automated Scoring: Homogeneous Groups

With regard to the RQ2, the 95% confidence intervals for the mean values provide a first indication of differences in the overall score between the responses of students, trainees, and in-service teachers. Looking at the students' responses, the automated rating using cross-validation in the population achieves an average PCK score between 5.44 and 6.13 (2011) and 4.76 and 5.49 points (2018) with a probability of 95%. The average PCK score using the student responses as training data is between 4.14 and 4.77 (2011) or 4.06 and 4.71 points (2018) with a probability of 95%. Since the two confidence intervals do not overlap in the respective samples, scoring engines using cross-validation and training data of student responses will most likely achieve a different overall

TABLE 4 | Item difficulty and discriminatory power (Kuhn, 2014).

Item	Difficulty	Corrected discriminatory power	Discriminatory power (ordered categories)	QWK H-M 2011 and 2018
A2	0.40	0.27	Score 0: −0.40 Score 1: 0.12 Score 2: 0.37	0.53
A4	0.45	0.33	Score 0: −0.55 Score 1: 0.12 Score 2: 0.47	0.59
A5	0.53	0.45	Score 0: −0.53 Score 1: 0.01 Score 2: 0.50	0.69
B10	0.66	0.38	Score 0: −0.36 Score 1: −0.25 Score 2: 0.48	0.55
C13	0.49	0.33	Score 0: −0.58 Score 1: 0.07 Score 2: 0.51	0.75
C15	0.39	0.33	Score 0: −0.54 Score 1: 0.17 Score 2: 0.45	0.73

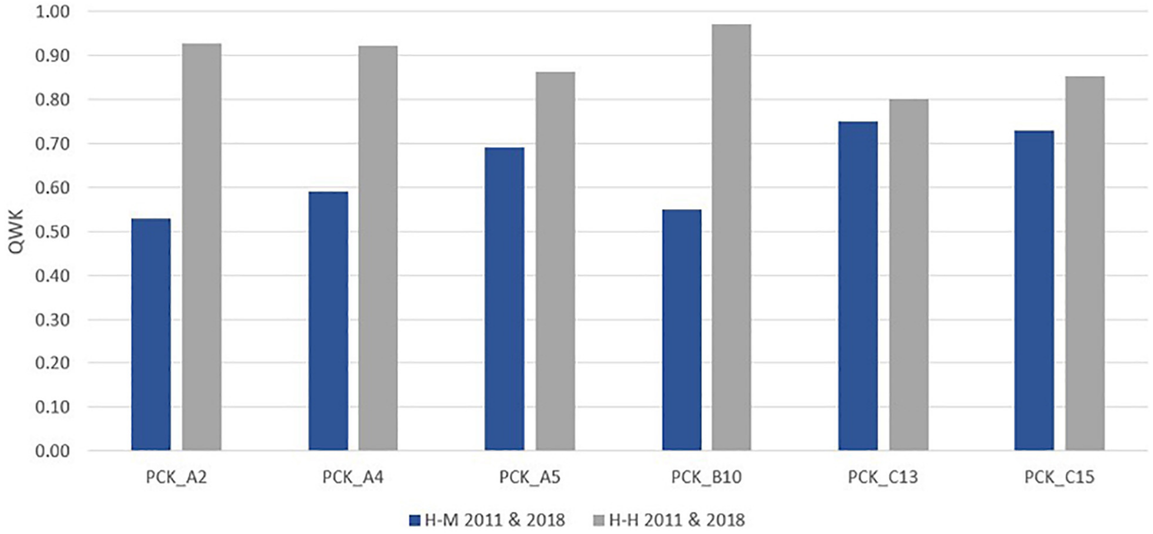


FIGURE 3 | Quadratically weighted kappa (QWK): Human-machine 2011 and 2018.

TABLE 5 | ANOVA[a] results on the variance factor of the homogeneous groups.

	Total PCK score		
	Cross validation vs. student answers	**Cross validation vs. answers of trainees**	**Cross validation vs. answers of in-service teachers**
2011	0.000*	0.003*	0.000*
2018	0.003*	0.000*	0.000*
2011/2018	0.608	0.000*	0.000*

[a] *Dependent variable: Total score of PCK in economics on open-ended items. *p < 0.05.*

PCK score. The 95% confidence intervals for the responses of trainees and in-service teachers support this assumption.

More precise results are provided by the significance values using ANOVAs (**Table 5**). The significance values ($0.000 < p < 0.003$) show differences within the comparison groups in all samples as well as in training models for the overall PCK score. Only the cross-validation and student response comparison in the cumulative sample (comprising the samples from 2011 to 2018) show a non-significance value ($p = 0.608$), which does not indicate any group difference. It can be assumed that training models in the population lead to different automated PCK scores when considering different groups.

The results of the variance analyses indicate that there is a difference in the overall PCK score when cross-validation and training data from homogeneous groups are used, and that this has an impact on the automated scoring results of PCK.

DISCUSSION AND CONCLUSION

This article deals with the question of the suitability of automated scoring for the rating of in-service teachers' PCK in the domain of economics. For this purpose, open-ended PCK tasks (Kuhn et al., 2016, 2020) were scored by human raters and the scoring engine ESCRITO (Zesch and Horbach, 2018). On average, there is a significant agreement between automated and human scoring. With regard to validity, the scoring results of computers and humans are convergent, suggesting that the automated scoring measures the same construct of PCK as the human scoring. Accordingly, the results of Kersting et al. (2014) can be confirmed. According to the current state of research, in teacher education, automated scoring is suitable for the purpose of supporting human raters in the evaluation of tasks such as those used in this PCK test to increase the validity and reliability of scoring while reducing the use of resources in assessment practice and research. However, automated scoring still has certain limitations, for example when it comes to the use of automated systems for more direct feedback in teacher education, for instance in the context of formative assessment.

Differences in automated scoring were particularly analyzed with reference to homogeneous groups. With regard to the training data used as a basis for the automated scoring, the results of the variance analyses show significant differences between cross-validation and homogeneous groups. This confirms the assumption that the use of training data consisting of responses of homogenous groups has an influence on automated scoring (Horbach and Zesch, 2019). To achieve scoring results that are as accurate as possible, the potential influence of the variance factor has to be taken into account when training scoring engines. Another possible influencing factor is the language in which the responses were formulated (Horbach and Zesch, 2019). Language skills were not analyzed in this paper, but gives cause for further investigation, since the PCK test used here has been adapted for the English-speaking region and has already been assessed (Brückner et al., 2017).

The number of responses, which comprise 5,112 texts, can be considered relatively low. More PCK responses and a larger amount of data would make it possible to train the automated scoring model to achieve a higher agreement between human and machine raters, especially for items A2, A4 and B10. A suitable automated scoring model (QWK above 0.70) thus allows for automated feedback systems to be used (Lee et al., 2019). At the item level, some items achieved considerably higher agreement values than others across all tests, indicating that the automated scoring differs between the PCK items and raising the question of whether and which item features contribute to this. Two possible features are the discriminatory power and item difficulty. However, the values (**Table 4**) did not indicate any correlation between the features and the human-machine agreement. Further analysis is required.

Another aspect associated with the rather small amount of data are the N-Gram models that were used. Based on these models, the probability that certain content (keywords and facts) is contained in the answers is calculated, which indicates a correct, partially correct or incorrect answer and leads to a corresponding scoring result. The question arises as to how exactly these probability calculations work with the available amount of data, and whether the calculations lead to word counting and whether the computer scoring measures an artificial construct (Perelman, 2014). This is not supported by the agreement results between human and computer, but those interrater reliabilities depend on the answers scored by human raters. Based on the scoring values, the engines learn and train the model used for the scoring. However, inconsistencies may occur in these human test scores (Bejar, 2012; Liu et al., 2014). Although inconsistencies in the PCK responses were reduced by data preparation, inconsistencies caused by the human raters (e.g., subjective perspectives, abilities, personality traits) are included in the learning process of the scoring engines, which can have an impact on automated scoring and the match between human and computer. To what extent such inconsistencies occur in the test scores given by human raters and to what extent they influence the automated PCK scoring remains to be investigated.

The use of further NLP concepts for the investigation of automated PCK scoring is also possible and should not be limited to the N-Gram model and vector regression. The three groups (students, trainees, and in-service teachers) show homogeneity with respect to the different stages of teacher education. Nevertheless, the groups are characterized by heterogeneity. For example, the students are divided into bachelor and master students. They also study at different universities. The group of trainees and in-service teachers also originates from different study seminars and schools, indicating that the participants in these groups had different learning opportunities (Horbach and Zesch, 2019). It is therefore questionable whether these three groups can be described as homogeneous and which criteria, in the sense of automated content analysis, a group has to meet to show homogeneity. In addition, it should be noted that the number of student responses (N) used for creating the automated scoring model were not very large. Therefore, the differences in the groups could also be due to the sample size. A larger amount of data would also be required for the purpose of invariance analyses.

With regard to the practical application in teacher education and research, automated scoring has the potential to support the human scoring. Automated scoring can reduce the amount of resources required for scoring and provide (prospective) teachers with more direct feedback (Öz and Özturan, 2018). Taking into account the abovementioned limitations and based on the moderate results, this paper provides a basis for formative assessment and automated feedback (Lee et al., 2019) on the facets of (prospective) teachers' professional competences, including PCK, using automated scoring and for integrating this technology into teacher training. To achieve this goal, however, further research is needed to achieve more substantial agreement between human raters and computers. Moreover, this paper only deals with one facet of teachers' professional competences. The automated scoring of further facets such as action-related and reflective competence (Kuhn et al., 2018, 2020) as well as the influence of further external factors (e.g., expert knowledge) on the scoring validity offers a broad field of further research possibilities.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the study participant.

AUTHOR CONTRIBUTIONS

AW planned and coordinated the study, and took the lead in writing the manuscript. The data collection is based on the preliminary work of CK and OZ-T. CK and OZ-T co-written and revised the manuscript. AW and CG jointly worked on scoring the data, analyzing, and interpreting the results. TZ and AH developed the Educational Scoring Toolkit. All authors contributed to the article and approved the submitted version.

REFERENCES

Almond, R. G. (2014). Using automated essay scores as an anchor when equating constructed response writing tests. *Int. J. Test.* 14, 73–91. doi: 10.1080/15305058.2013.816309

Alonzo, A. C., Kobarg, M., and Seidel, T. (2012). Pedagogical content knowledge as reflected in teacher-student interactions: analysis of two video cases. *J. Res. Sci. Teach.* 49, 1211–1239. doi: 10.1002/tea.21055

Attali, Y. (2013). "Validity and reliability of automated essay scoring," in *Handbook of Automated Essay Evaluation*, eds M. D. Shermis and J. Burstein (New York, NY: Routledge), 181–198.

Basu, S., Jacobs, C., and Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Trans. Assoc. Comput. Ling.* 1, 391–402. doi: 10.1162/tacl_a_00236

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educ. Measur. Issues Pract.* 31, 2–9. doi: 10.1111/j.1745-3992.2012.00238.x

Bejar, I. I., Mislevy, R. J., and Zhang, M. (2016). "Automated scoring with validity in mind," in *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, eds A. A. Rupp and J. P. Leighton (Oxford: Wiley-Blackwell), 226–246.

Bridgeman, B., Trapani, C., and Attali, Y. (2012). Comparison of human and machine scoring of essays: differences by gender, ethnicity, and country. *Appl. Measur. Educ.* 25, 27–40. doi: 10.1080/08957347.2012.635502

Brückner, S., Kuhn, C., Day, S., Zlatkin-Troitschanskaia, O., and Saas, H. (2017). "Adapting a german test instrument to evaluate subject-specific instructional skills of teachers of economics in the USA," in *Presentation at the 56th Annual Financial Literacy and Economic Education Conference*, (New York, NY: Counsil for Economic Education).

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *Int. Artif. Intell. Educ.* 25, 60–117. doi: 10.1007/s40593-014-0026-8

Burstein, J., Tetreault, J., and Madnani, N. (2013). "The e-rater® automated essay scoring system," in *Handbook of Automated Essay Evaluation*, eds M. D. Shermis and J. Burstein (New York, NY: Routledge), 55–67.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: fishing for red herrings? *Assess. Writing* 18, 100–108. doi: 10.1016/j.asw.2012.11.001

Daxenberger, J., Ferschke, O., Gurevych, I., and Zesch, T. (2014). "DKPro TC: a java-based framework for supervised learning experiments on textual data," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Balitmor, MD: Association for Computational Linguistics), 61–66. doi: 10.3115/v1/P14-5011

Dolan, R., and Burling, K. (2012). ""Computer-based testing in higher education"," in *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, eds C. Secolsky and D. B. Denison (New York, NY: Routledge), 341–355.

Drolia, S., Rupani, S., Agarwal, P., and Singh, A. (2017). Automated essay rater using natural language processing. *Int. J. Comput. Appl.* 163, 44–46. doi: 10.5120/ijca2017913766

Evanini, K., Heilman, M., Wang, X., and Blanchard, D. (2015). *Automated Scoring for the TOEFL Junior® Comprehensive Writing and Speaking Test. Research Report ETS RR-15-09*. Avaialable online at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ets2.12052 (accessed March 17, 2020).

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0471445428

Gerard, L. F., and Linn, M. C. (2016). Using automated scores of student essays to support teachr guidance in classroom inquiry. *J. Sci. Teach. Educ.* 27, 111–129. doi: 10.1007/s10972-016-9455-6

Heilman, M., and Madnani, N. (2015). "The impact of training data on automated short answer scoring performance," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, (Denver, Co: Association for Computational Linguistics), 81–85.

Horbach, A., Scholten-Akoun, D., Ding, Y., and Zesch, T. (2017). "Fine-grained essay scoring of a complex writing task for native speakers," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, (Copenhagen: Association for Computational Linguistics), 357–366.

Horbach, A., Stennmanns, S., and Zesch, T. (2018). "Cross-lingual content scoring," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, (New Orleans, LA: Association for Computational Linguistics), 410–418.

Horbach, A., and Zesch, T. (2019). The influence of variance in learner answers on automatic content scoring. *Front. Educ.* 4:28. doi: 10.3389/feduc.2019.00028

Kelava, A., and Moosbrugger, H. (2012). "Deskritpivstatistische evaluation von Items (Itemanalyse) und testwertverteilung," in *Testtheorie und Fragebogenkonstruktion*, eds H. Moosbrugger and A. Kevala (Berlin: Srpinger VS), 75–102. doi: 10.1007/978-3-642-20072-4

Kersting, N. B., Sherin, B. L., and Stigler, J. W. (2014). Automated scoring of teachers' open-ended responses to video prompts: bringing the classroom-video-analysis assessment to scale. *Educ. Psychol. Measur.* 74, 950–974.

Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-Verwaltenden Bereich*. Landau: Verlag Empirische Pädagogik.

Kuhn, C., Alonzo, A. C., and Zlatkin-Troitschanskaia, O. (2016). Evaluating the pedagogical content knowledge of pre- and in-service teachers of business and economics to ensure quality of classroom practice in vocational education and training. *Empirical Res. Voc. Ed. Train.* 8:5. doi: 10.1186/s40461-016-0031-2

Kuhn, C., Zlatkin-Troitschanskaia, O., Brückner, S., and Saas, H. (2018). A new video-based tool to enhance teaching economics. *Int. Rev. Econ. Educ.* 27, 24–33. doi: 10.1016/j.iree.2018.01.007

Kuhn, C., Zlatkin-Troitschanskaia, O., Lindmeier, A., Jeschke, C., Saas, H., and Heinze, A. (2020). "Relationships between domain-specific knowledge, generic attributes, and instructional skills – results from a comparative study with pre- and in-service teachers of mathematics and economics," in *Student Learning in German higher Education: Innovative Measurement Approaches and Research Results*, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Wiesbaden: Springer VS), 75–103. doi: 10.1007/978-3-658-27886-1_5

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

Leacock, C., and Chodorow, M. (2003). C-rater: automated scoring of short-answer questions. *Comput. Hum.* 37, 389–405.

Lee, H., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., and Liu, O. L. (2019). Automated text scoring and real—time adjustable feedback: supporting revision of scientific arguments involving uncertainty. *Sci. Educ.* 103, 590–622. doi: 10.1002/sce.21504

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., and Linn, M. C. (2014). Automated scoring of constructed-response science items: prospects and obstacles. *Educ. Measur. Issues Pract.* 33, 19–28.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., and Linn, M. C. (2016). Validation of automated scoring of science assessments. *J. Res. Sci. Teach.* 53, 215–233.

Loukina, A., and Cahill, A. (2016). "Automated scoring across different modalities," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, (San Diego, CA: Association for Computational Linguistics), 130–135.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. Med.* 22, 276–282.

Öz, H., and Özturan, T. (2018). Computer-based and paper-based testing: does the test administration mode influence the reliability and validity of achievement tests? *J. Lang. Ling. Stud.* 14, 67–85.

Page, E. B. (1966). The imminence of Grading essays by computer. *Phi Delta kappa* 47, 238–243.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing* 21, 104–111.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., and Bridgeman, B. (2012a). *Evaluation of the e-rater® Scoring Engine for the GRE® Issue and Argument Prompts. Research Report ETS RR–12-02*. Available online at: https://www.ets.org/Media/Research/pdf/RR-12-02.pdf (accessed March 17, 2020).

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., and Bridgeman, B. (2012b). *Evaluation of the e-rater® Scoring Engine for the TOEFL® Independent and Integrated Prompts. Research report ETS RR–12-06*. Available online at: https://www.ets.org/Media/Research/pdf/RR-12-06.pdf (accessed March 17, 2020).

Riordan, B., Horbach, A., Cahill, A., Zesch, T., and Lee, C. M. (2017). "Investigating neural architectures for short answer scoring," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, (Copenhagen: Association for Computational Linguistics), 159–168. doi: 10.18653/v1/W17-5017

Shavelson, R. J. (2009). *Measuring College Learning Responsibly*. Stanford, CA: Stanford University Press.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: competition, results, and future directions from a United States demonstration. *Assess. Writing* 20, 53–76. doi: 10.1016/j.asw.2013.04.001

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educ. Assess.* 20, 46–65. doi: 10.1080/10627197.2015.997617

Shermis, M. D., Burstein, J., and Bursky, S. A. (2013). "Introduction to automated essay evaluation," in *Handbook of Automated Essay Evaluation*, eds M. D. Shermis and J. Burstein (New York: Routledge), 1–15.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14.

Sim, J., and Wright, C. G. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. Ther.* 85, 257–268.

Stuhlsatz, M., Wilson, C. D., Buck-Bracey, Z., Urban-Lurain, M., Merrill, J. E., and Haudek, K. C. (2016). "Applying automated analysis to develop a cost-effective measure of science teacher pedagogical content knowledge," in *Paper Presented at National Assocation for Research in Science Teaching Annual Conference*, (Baltimore, MD: NARST).

Wilson, C., Stuhlsatz, M., Bracey, Z. B., and Donovan, B. (2017). "Applying automated analysis to the measurement of constructed responses: applications in student argumentation and teacher PCK," in *Paper Presented at the European Science Education Research Association Conference*, (Dublin: Dublin City University),

Wilson, C. D., Stuhlsatz, M., Hvidsten, C., and Gardner, A. (2018). "Analysis of practice and teacher PCK: inferences from professional development research," in *Pedagogical Content Knowledge in STEM*, eds S. M. Uzzo, S. B. Graves, E. Shay, M. Harford, and R. Thompson (Cham: Springer), 3–16.

Wind, S. A., Wolfe, E. W., Engelhard, G. Jr., Foltz, P., and Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *Int. J.0 Test.* 18, 27–49. doi: 10.1080/15305058.2017.1361426

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G., and Cunningham, S. J. (1999). *Weka: Practical Machine Learning Tools and Techniques with Java Implementations. (Working paper 99/11)*. Hamilton: University of Waikato.

Zehner, F., Goldhammer, F., and Sälzer, C. (2018). Automatically analyzing text responses for exploring gender-specific cognitions in PISA reading. *Large Scale Assess. Educ.* 6:7. doi: 10.1186/s40536-018-0060-3

Zehner, F., Sälzer, C., and Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educ. Psychol. Measur.* 76, 280–303. doi: 10.1177/0013164415590022

Zesch, T., and Horbach, A. (2018). "ESCRITO – an NLP-enhanced educational scoring toolkit," in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, (Miyazaki: European Language Resources Association (ELRA), 2310–2316.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). "Task-independent features for automated essay grading," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, (Denver, CO: Association for Computational Linguistics), 224–232.

Zhang, M. (2013). *Contrasting Automated and Human Scoring of Essays*. Princeton, NJ: Educational Testing Services.

Zlatkin-Troitschanskaia, O., Kuhn, C., Brückner, S., and Leighton, P. J. (2019). Evaluating a technology-based assessment (TBA) to measure teachers' action-related and reflective skills. *Int. J. Test.* 19, 148–171. doi: 10.1080/15305058.2019.1586377

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.