Bringing Automatic Scoring into the Classroom – Measuring the Impact of Automated Analytic Feedback on Student Writing Performance

Andrea Horbach[1], Ronja Laarmann-Quante[2], Lucas Liebenow[3], Thorben Jansen[3], Stefan Keller[4], Jennifer Meyer[3], Torsten Zesch[1] and Johanna Fleckenstein[3,5]
[1]CATALPA, FernUniversität in Hagen, Germany, [2]Ruhr-Universität Bochum, Germany,
[3]Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel, Germany,
[4]Pädagogische Hochschule Zürich, Switzerland, [5]Universität Hildesheim, Germany

Abstract

While many methods for automatically scoring student writings have been proposed, few studies have inquired whether such scores constitute effective feedback improving learners' writing quality. In this paper, we use an EFL email dataset annotated according to five analytic assessment criteria to train a classifier for each criterion, reaching human-machine agreement values (kappa) between .35 and .87. We then perform an intervention study with 112 lower secondary students in which participants in the feedback condition received stepwise automatic feedback for each criterion while students in the control group received only a description of the respective scoring criterion. We manually and automatically score the resulting revisions to measure the effect of automated feedback and find that students in the feedback condition improved more than in the control group for 2 out of 5 criteria. Our results are encouraging as they show that even imperfect automated feedback can be successfully used in the classroom.

1 Introduction

Writing e-mails in English is an important skill in many academic and professional contexts and, thus, part of many secondary school curricula in English as a foreign language (EFL). However, scoring writing exercises manually and providing feedback is a time-consuming task for educators. Therefore, we present a study on how to automatically provide feedback based on automated scores. The study took place in the context of EFL education at secondary level in Switzerland and Germany. In contrast to other studies that focus only on the technical evaluation of a machine learning approach, we go one step further and directly measure the effects of using automatic scoring to provide feedback in the classroom. We conducted this experiment as a controlled randomized experimental study.

To this end, we first describe the dataset this study is based on. The eRubrix corpus (Keller et al., 2023) contains a total of 1,104 semi-formal e-mails written in response to three different prompts (see below for details).In these e-mail texts, five individual trait scores are annotated, assessing whether individual parts of an e-mail are addressed in an appropriate fashion. Table 1 shows an example from the dataset: the original draft as well as the five revisions produced by a participant in the feedback group.

We then describe an NLP pipeline used to automatically score this dataset analytically according to these five criteria. Besides the prompt-specific scoring used in our intervention study, we also provide additional experiments evaluating cross-prompt scoring performance in order to show the transferability of the approach to new writing prompts of a similar kind. In the subsequent experimental study, we show the usefulness of feedback generated from the automatic score, comparing the performance improvement of an intervention group (receiving informative tutorial feedback) with that of a control group (receiving scoring criteria only). In this study, we show that students in the feedback group improved more than students in the control group for two out of five criteria.

2 Related Work

In this section, we first contextualize our scoring task within the automatic scoring landscape and then introduce the psychological background of our intervention study.

Automatic Scoring The task tackled in this paper is an instance of essay scoring in which we assess texts both according to their linguistic quality and their content (Beigman Klebanov and Madnani, 2020). The setup in which different aspects of an essay are scored is similar to what is of-

Andrea Horbach, Ronja Laarmann-Quante, Lucas Liebenow, Thorben Jansen, Stefan Keller, Jennifer Meyer, Torsten Zesch and Johanna Fleckenstein. Bringing Automatic Scoring into the Classroom - Measuring the Impact of Automated Analytic Feedback on Student Writing Performance. *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022)*. Linköping Electronic Conference Proceedings 190: 72–83.

E-mail Text	Criterion	Score
English questions Hello, Is a three- week course possible? I think two weeks courses for all levels, qualified from experienced teachers. And how much is the price? Who organized the activities and what of activities are organized? See you Kim Weber	Content Completeness	Pass
English **learning** Hello, Is a three- week course possible? I think two weeks courses for all levels, qualified from experienced teachers. **Could you tell me** how much is the price? **And** who organized the activities and what of activities are organized? See you Kim Weber	Greeting & Closing	Fail
English learning **Dear Mrs Black,** Is a three- week course possible? I think two weeks courses for all levels, qualified from experienced teachers. Could you tell me how much is the price? And who organized the activities and what of activities are organized? **Best wishes** Kim Weber	Subject Line	Pass
Questions at the Central School Dear Mrs Black, Is a three- week course possible? I think two weeks courses for all levels, qualified from experienced teachers. Could you tell me how much is the price? And who organized the activities and what of activities are organized? Best wishes Kim Weber	Interpersonal Dimension	Fail
Questions at the Central School Dear Mrs Black, **I'm writing to tell you my questions and I would like to ask you about the Central School.** Is a three- week course possible? I think two weeks courses for all levels, qualified from experienced teachers. Could you tell me how much is the price? And who organized the activities and what of activities are organized? **Thank you for answering my questions.** Best wishes Kim Weber	Register & Style	Fail
Questions at the Central School Dear Mrs Black, I'm writing to tell you my questions and I would like to ask you about the Central School. Is a three- week course possible? I think two weeks courses for all levels, qualified from experienced teachers. Could you tell me how much is the price? **Finally** ,who organized the activities and what of activities are organized? Thank you for answering my questions. Best wishes Kim Weber	Final Revision	-

Table 1: An example e-mail written in response to the 'Language School' prompt in the eRubrix dataset. We show the original e-mail together with its five revisions (edits are highlighted by the authors) and whether the e-mail passed or failed the respective criterion.

ten called trait-based essay scoring (Lee et al., 2010). However, an important difference is that in most work, essay traits are considered one dimension according to which to score a *whole text*, such as coherence (Yannakoudakis and Briscoe, 2012; Farag et al., 2018), topicality (Klebanov et al., 2016) or argumentation (Stab and Gurevych, 2014; Persing and Ng, 2015, 2016). In contrast, human judgments for each rubric in the eRubrix dataset only refer to *specific parts* of an essay and score them according to their appropriateness. This is similar to a holistic score for only a subpart of the text as in Horbach et al. (2017), where essays consist of a summary and a discussion part scored separately. (Note that in our automatic scoring, we nevertheless use the whole text as input in most cases, as we cannot reliably split the data into individual segments). This makes it similar to the task of facet-based short-answer scoring (Nielsen et al., 2009, 2008) where the presence of certain content units, so called facets, in the text is analyzed. However, one crucial difference is that in our case both content and form are scored together.

Further, the task of writing an e-mail or letter is well known in automated essay scoring. The ASAP-AES dataset, for example, also contains tasks where students have to write a letter.[1] However, such tasks are often framed in terms of a persuasive text that conveys the author's own position, whereas in our task, e-mails are written in order to gather information.

Feedback Intervention Study The aim of our intervention study is to investigate the effect of informative tutorial feedback based on automatically scored texts. In instructional contexts, feedback generally refers to any information given to a person during or after a learning process. It aims to reduce the gap between the current performance and the desired learning outcome (Mory, 2004; Narciss, 2008; Sadler, 1989). Feedback is deemed one of the most effective factors influencing student learning, however, meta-analyses show that the effects are heterogeneous (for feedback on learning in general: cf. Wisniewski et al., 2020; for feedback on writing: cf. Graham et al., 2015). Attempting to explain the inconsistent findings, certain moderators for feedback effectiveness have been identified (Bangert-Drowns et al., 1991; Black and Wiliam, 1998; Hattie and Timperley, 2007; Kluger and DeNisi, 1996; Mory, 2004; Shute, 2008). Feedback has a positive effect on learner performance only if it reduces uncertainty and cognitive load by presenting the information necessary to improve task performance. According to Narciss (2008), informative tutorial feedback should include both evaluative information (i.e., information on the current task performance) and tutorial information (i.e., elaborate information to improve task performance) in order to support learning effectively. Hattie and Timperley's 2007 feedback model summarizes the empirically identified effectiveness criteria using three questions: "Where am I going?" (transparency of learning goals), "How am I going?" (individual information on current task performance), and "Where to next?" (information on how to achieve learning goals).

In accordance with this model, feedback was conceptualized according to these criteria in our study. Learners were presented with evaluative information on their performance (aspect mastered/not mastered) as well as elaborative feedback (hints and examples for performance improvement).

The evidence on the effectiveness of automatic feedback on writing performance is also described as being heterogeneous (McNamara et al., 2015; Stevenson and Phakiti, 2014; Strobl et al., 2019). Fleckenstein et al. (in press) conducted a systematic review of individual writing support by intelligent tutoring systems (ITS). Whereas the effects of the interventions were promising in general, the authors found that there were only few studies with randomized controlled experimental designs (see, e.g., Kellogg et al., 2010; Palermo and Thomson, 2018; Wade-Stein and Kintsch, 2004; Wilson and Roscoe, 2020; Wilson and Czik, 2016; Xu and Zhang, 2022). Moreover, it was often unclear what type of tutorial support led to performance improvement as the interventions often included non-adaptive, confounding support measures (e.g., pre-writing activities, strategy instruction, drill and practice) in addition to holistic and/or analytic automated feedback. Our intervention study is one of the few randomized controlled experiments that investigates the unconfounded effect of analytic feedback in the context of automated scoring.

[1] https://www.kaggle.com/c/asap-aes

Prompt	# e-mails	∅ # tokens (SD)
Language School	368	97.9 (± 33.0)
Burger Restaurant	369	104.1 (± 34.0)
Camping	367	105.0 (± 34.1)

Table 2: Basic dataset statistics.

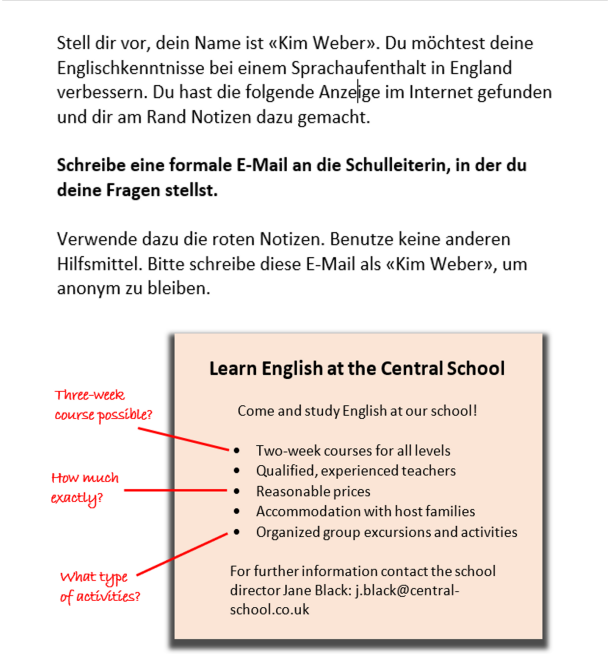


Figure 1: Instructions for the *language school* prompt. The German text translates as follows: *Imagine your name is Kim Weber. You want to improve your English language skills through a language stay in England. You have seen the following ad on the Internet. Write a formal e-mail to the school principal asking your questions. Use the notes printed in red. Do not use any other material. Write the e-mail as 'Kim Weber' to stay anonymous.*

3 Data

The eRubrix dataset contains three individual writing prompts, each asking the student to write an information-seeking e-mail. In the first task, students inquire about attending a course at a language school in the UK, in the second task, they respond to a job advertisement at a burger restaurant, and in the third, they gather information for a camping holiday.

One is an inquiry Table 2 shows basic statistics for the dataset. Figure 1 shows as an example of the language school prompt. Per prompt, about 370 individual e-mails were collected.

Each e-mail was scored with a binary label for each of the following criteria, corresponding to key elements of an e-mail. The description of each criterion closely follows the scoring rubrics described in Keller et al. (2023).

- **Content Completeness:** whether the e-mail asks for all three pieces of information required in the task.
- **Greeting & Closing:** whether the salutation at the beginning and the closing are adequate to the situation.
- **Subject Line:** whether the subject line adequately communicates the intention of the e-mail.
- **Interpersonal Dimension:** whether writers explain who they are, what the purpose of the mail is and describe at the end what kind of response they expect.
- **Register & Style:** whether the e-mail uses clear, detailed and adequate language and is free from mistakes which inhibit understanding.

Scoring was performed by two trained annotators, cases of disagreement were adjudicated by a third annotator. Table 3 shows inter-annotator-agreement (Cohen's kappa), as well as the label distribution by indicating the fraction of texts that mastered the respective criterion. We see that annotators were able to agree on the first four criteria well, while *Register & Style* seemed to be more problematic to annotate.

4 Automatic Scoring

In this section, we describe our automatic scoring procedure. After the experimental setup, we report experiments for prompt-specific scoring where one classifier is trained per prompt and per scoring rubric We also perform generic scoring with a model trained across prompts, i.e. on more training data. The prompt-specific model for the *language school* prompt is used in our intervention study. In order to show the transferability of our approach, we also report on additional experiments for cross-prompt training.

4.1 Experimental Setup

We use the Gradient Boosting classifier from scikit-learn[2] with a maximum tree depth of 6

[2]https://scikit-learn.org

Prompt	Content		Greeting		Subject		Interpersonal		Style	
	% corr.	IAA	% corr.	IAA	% corr.	IAA	% corr.	IAA	% corr.	IAA
All	87.6	-	31.2	-	72.6	-	62.0	-	19.7	-
Language School	87.5	.88	26.4	.90	67.9	.68	59.0	.91	22.0	.43
Burger Restaurant	87.5	.80	36.3	.93	89.5	.96	62.3	.94	19.5	.38
Camping	87.7	.85	30.8	.89	60.8	.75	64.9	.89	17.7	.47

Table 3: Label distribution (%corr. marks the percentage of essays where the criterion was fulfilled) and inter-annotator agreement for each scoring rubric, measured in Cohen's kappa.

Train	Test	Content		Greeting		Subject		Interpersonal		Style	
		acc	κ	acc	κ	acc	κ	acc	κ	acc	κ
All (CV)		.93	.59	.89	.75	.95	.88	.88	.75	.84	.38
Language School (CV)		.92	.60	.88	.67	.94	.87	.85	.69	.81	.35
Burger Restaurant (CV)		.94	.69	.92	.83	.99	.96	.82	.60	.82	.29
Camping (CV)		.91	.51	.89	.73	.93	.86	.77	.47	.82	.26
Burger & Camping	School	.83	.38	.76	.30	.81	.62	.89	.78	.68	.20
School & Camping	Burger	.93	.66	.85	.64	.67	.25	.85	.69	.82	.33
School & Burger	Camping	.50	.10	.75	.33	.71	.46	.84	.66	.85	.39

Table 4: Experimental results measured in accuracy and Cohen's kappa for cross-validation experiments on all data and per prompt (upper half) as well as prompt transfer between prompts (lower half).

and otherwise standard parameters and TF-IDF weighted unigram features. We evaluate using accuracy as well as Cohen's kappa (Cohen, 1960) as a way of measuring chance-corrected agreement.

4.2 Prompt-specific vs Generic Scoring

In a first set of experiments, we compare two different setups. We either train a generic model using data from all three prompts as training material or we train a prompt-specific model using only data from the same prompt for training and testing. In other words, we compare whether a model benefits from more training data coming from a different prompt. In both setups, we use ten-fold cross-validation.

The upper half of Table 4 shows the results. We see that we get a slight advantage for the two categories *Interpersonal* and *Style* when using more training data from other prompts, whereas this is only partially helpful for the other three criteria (*Content*, *Greeting* and *Subject*). We speculate that this is because the latter three are the most content dependent and therefore mainly rely on the specific lexical material for one prompt, while the other two contain also generic lexical material, like, e.g., "I am looking forward to your answer". Generally, we see that the highest prediction performance can be achieved for the *Subject* line, while *Style* is hardest to predict, which is probably due to the high class imbalance of this criterion, i.e., there are only few instances in the training data where the criterion is mastered. This criterion is also difficult to score for human raters as evidenced by the agreement scores, which are much lower than for the other criteria.

4.3 Cross-prompt Scoring

In order to asses the usability of the models in a real-life scenario where training data for a new prompt might not be readily available, we investigate model transfer to new prompts not used during training.

To do so, we train on all data from two prompts and test on the third prompt. The results are shown in the lower part of Table 4. We see that for most rubrics, the performance drops considerably compared to the within-prompt setting. However, for *Interpersonal* and *Style*, we partially find an improvement of the results in the cross-prompt setting. We assume that similar to our finding for the *All* setting above, these two rubrics rely a lot on generic wording. In addition, the Style rubric has a high class imbalance for the *Camping* setting, which might explain why this prompt is particularly susceptible to cross-prompt (and more balanced) training data.

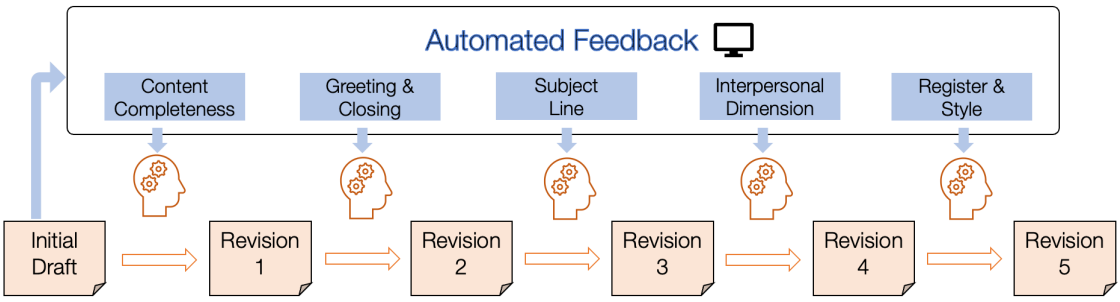


Figure 2: Students sequentially receive automated feedback on their original e-mail and are given the opportunity to revise based on the feedback.

5 Feedback Intervention Study

We investigated the following research questions: (1) Does the automated feedback lead to substantial improvement in students' writing performance? (2) What do we learn about the revision process by looking at stepwise text development? In the following, we first describe the procedure and results of our intervention study and then provide further insights into the resulting e-mail revision dataset.

5.1 Procedure

We conducted a randomized controlled field experiment with $N = 112$ lower secondary (ISCED level 2[3]) students to investigate the effect of a feedback intervention that was based on the automatic assessment. Seven students were excluded from the sample due to incomplete data, leaving a final sample of N = 105 students (n = 53 female; age M = 14.41, SD = 0.81) in grade 8 (n = 54) and grade 9 (n = 51) for the statistical analyses. Students were asked to respond to the 'language school' e-mail writing prompt (see Figure 1) that was then assessed using the scoring model for that specific prompt as described above.

As part of the intervention, students received automatic feedback on the five assessment criteria and were asked to revise their text accordingly. To communicate the feedback in the process of writing, a scoring rubric was used which contained the most important elements of the genre 'e-mail' (Keller et al., 2023). The elements were arranged in a stepwise manner based on the principle of communicative effectiveness (Widdowson, 1978), and presented to students in sequential order so that they could focus on one criterion at a time before moving on to the next one. In a process- and genre-based approach to writing (Hyland, 2007), feedback guided students towards writing good e-mails by focusing their attention on important generic elements by the principle of increasing communicative value.

Within the writing tasks set in this study, the most important element was to include all the questions mentioned in the task. Therefore, this element appeared as Step 1 in the rubric. If texts were found to be lacking one or several questions, the feedback suggested to go back to the task and make sure they had covered all required aspects. In subsequent steps, students were advised to contextualize their e-mails by finding appropriate formulas of salutation and closing (Step 2), to formulate clear and precise subject lines (Step 3), and to frame their e-mails with an introduction stating their name and the nature of their inquiry, and an indication of what type of answer they expected (Step 4). Finally, students were advised to check the grammar, lexis and spelling of their e-mail to make sure it did not contain any formal mistakes.

The decision to place formal correctness (*Register & Style*) as the last step in the feedback process was based on the assumption that it is easier for learners to master the specific elements of a genre (which can be explicitly taught and learned) than to make progress in the general aspects of foreign language proficiency, such as syntax or lexical quality. Further, focusing their attention on formal mistakes too early would have risked students getting bogged down with questions of linguistic correctness, while the focus of the intervention lay on using language in a communicative way (Keller et al., 2023). Figure 2 visualizes the revision process.

Students were randomly assigned to the feed-

[3] https://iqa.international/isced-levels/

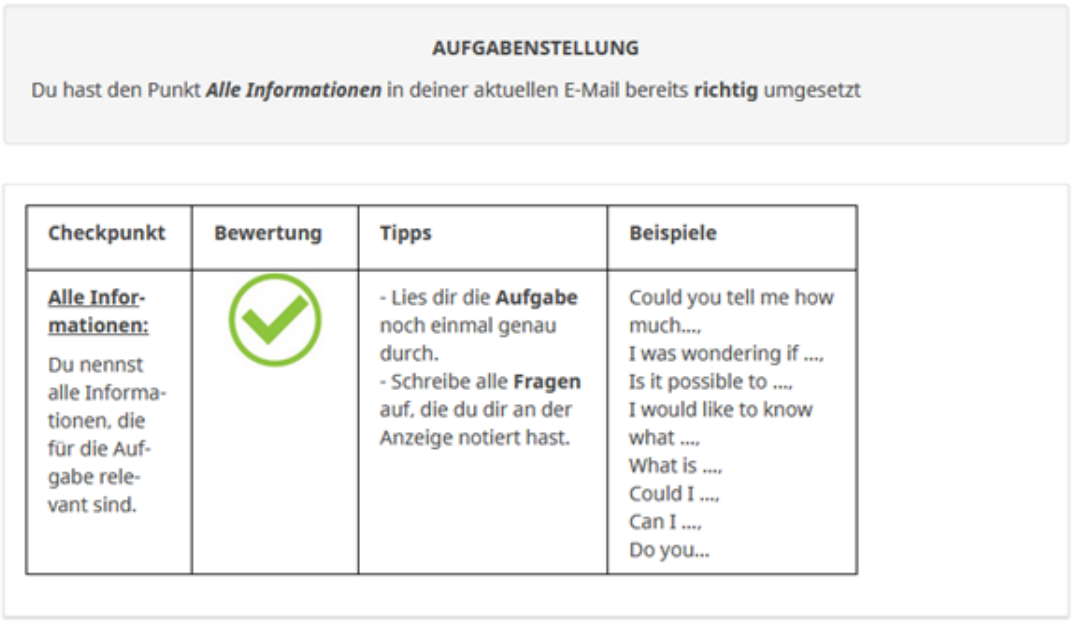


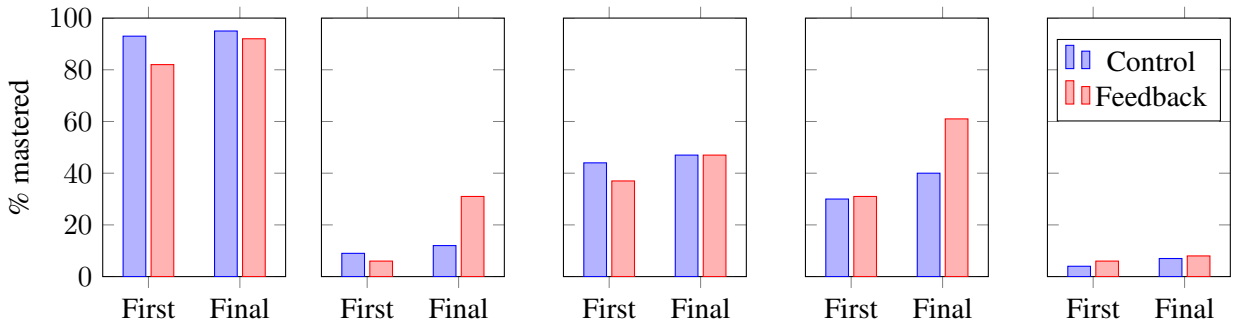
Figure 3: Example for a feedback message received in the category *Content*. Students from the control group were shown the requirements only (left column), while students in the feedback condition received their automatic score together with hints how to improve their writing.

back condition or the control condition. The feedback group was provided with informative tutorial feedback in German, including both evaluative and elaborative information on each scoring criterion including exemplary formulations in English. See Figure 3 for an example for the criterion *Interpersonal Dimension*. The first column specifies the requirements (e.g., 'Do you explain who you are and why you are writing?') for passing that criterion, the second one visualizes the predicted score. The third column contains hints how to improve the writing (e.g., 'Introduce yourself in the first sentence') while the fourth column contains concrete examples of appropriate formulations. The control group was provided with a description of the scoring criteria (i.e. only the first column in Figure 3), but did not receive individual feedback on their performance. All texts were scored on the five assessment criteria, using binary codes: 0 = *criterion not mastered* and 1 = *criterion mastered*.

We compared the performance on each criterion between the two groups before and after the feedback intervention, expecting the feedback group to show more substantial improvement. As the outcome was dichotomous (0/1), we analyzed the data using the R package nparLD (Noguchi et al., 2012), which allows for the nonparametric analysis of longitudinal data in factorial experiments. Wald-tests (Wald, 1943) were performed to test whether the interaction of group (control vs. feedback) and time (initial draft vs. final draft) was statistically significant for each of the five criteria.

5.2 Results

Performance Improvement Figure 4 shows the performance results based on the automatic scoring for the two groups on the first draft and on the final revised version. For each criterion, the graphs show what proportion of students had successfully mastered the criterion. For *content completeness*, the vast majority of students in both the control group (93 %) and the feedback group (82 %) had already met the requirement in their first draft. In the feedback group, 10 percent were able to improve in the revision whereas the control group remained at a consistently high level (95 %). The group differences in content completeness improvement, however, were not statistically significant, χ^2=3.22; *ns*. Only a small minority of the students mastered the criterion *Greeting and Closing* in their first draft (9 % in the control, 6 % in the feedback condition), showing little improvement in the control group (12 %) and substantial improvement in the feedback group (31 %). This difference in improvement between the groups was statistically significant, χ^2=9.88; p<.01. The criterion *subject line* was fulfilled by 44 percent (con-



(a) Content (b) Greeting (c) Subject Line (d) Interpersonal (e) Register/Style

Figure 4: Improvements for control group and feedback group according to the automatic scoring model.

trol) and 37 percent (feedback), respectively, before the intervention, and by 47 percent in both groups after the intervention. While the descriptive results suggest that the feedback group was able to catch up, the effect was not significant, χ^2=1.19; *ns*. In both groups, almost a third of the students (30 % in the control, 31 % in the feedback condition) already mastered the criterion *interpersonal dimension* before the intervention. This percentage increased to 61 percent in the feedback condition and only 40 percent in the control condition. This effect was significant, χ^2=6.61; p <.01. The criterion *register and style* was only met by very few students before (4 % in the control, 6 % in the feedback condition) and after (7 % in the control, 8 % in the feedback condition) the intervention, yielding no significant differences, χ^2=0.29; *ns*.

5.3 Follow-up Analyses

The experiment resulted in an e-mail revision dataset where 5 revisions for each e-mail were recorded. This offers a unique opportunity to get insights into the properties of these revisions as well as the scoring behaviour of the trained model under realistic conditions.

E-mail Length As a first proxy for the extent of revisions we tracked e-mail length across revisions. Figure 5 shows the number of characters for each revision step in each condition.

We can see that there is a large variance of e-mail lengths at all revision steps, especially for the control group. In both groups, there is only a slight tendency that e-mails get longer across revisions, which indicates that the students do not primarily revise their texts by adding more content.

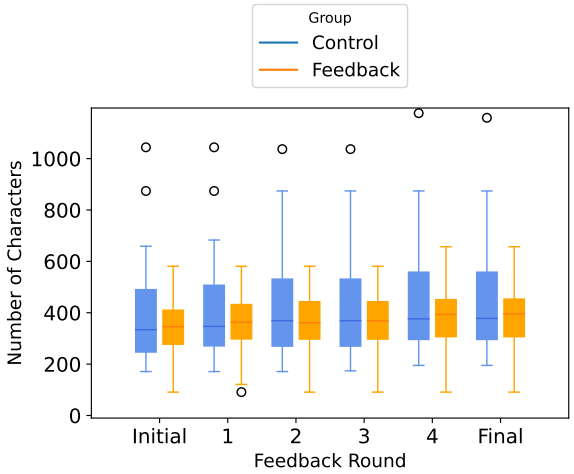


Figure 5: E-mail length (measured in characters) at each revision step.

Extent of Revisions To further investigate the nature of the revisions, we compute character-based edit distance between subsequent revisions, i.e. we count the minimal number of insertions, deletions or substitutions from one version to the next revision for both the feedback and the control condition.

Figure 6 shows that both groups display a similar pattern with most edits done after the initial step and the third revision. Manual inspection of essays from both groups showed that, in the first revision, students sometimes completed a not yet finished e-mail.

For the feedback group, we further looked separately at those students who were given the feedback that they had already mastered a criterion in contrast to those who were given the information that the criterion was not yet mastered. Figure 7 reveals that after the initial review, only those students which had not yet mastered a criterion made any revisions to their texts.

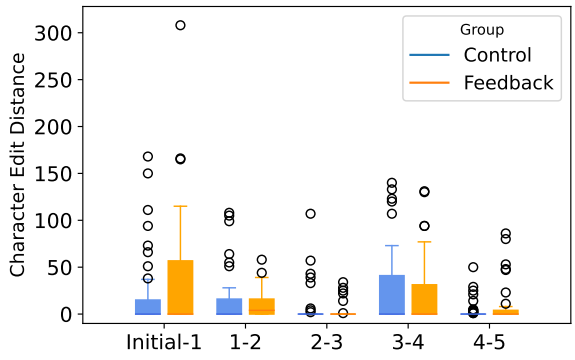


Figure 6: Edit distance between two consecutive revision steps (i.e. 2-3 is the edit distance between revision 2 and 3)

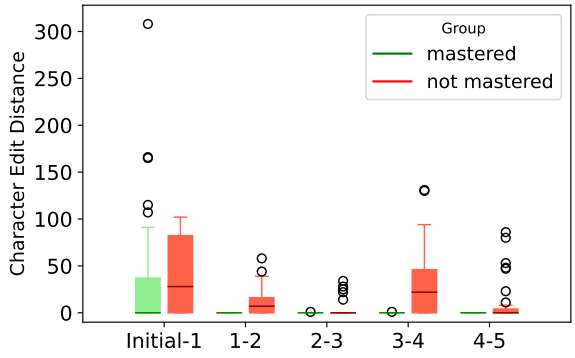


Figure 7: Edit distance between two consecutive revision steps (i.e. 2-3 is the edit distance between revision 2 and 3) for the Feedback group divided into those who passed or failed a certain criterion.

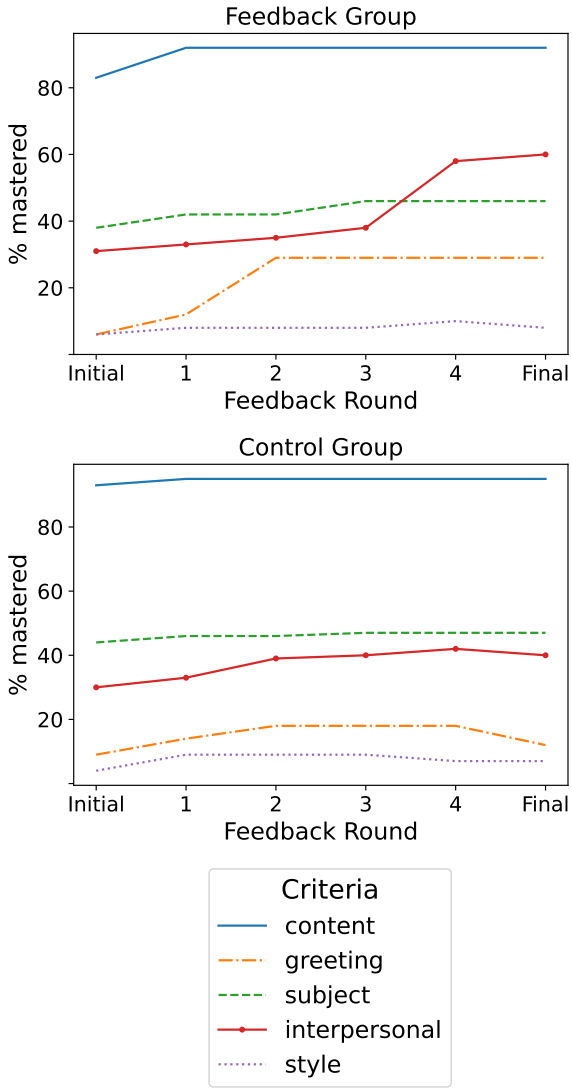


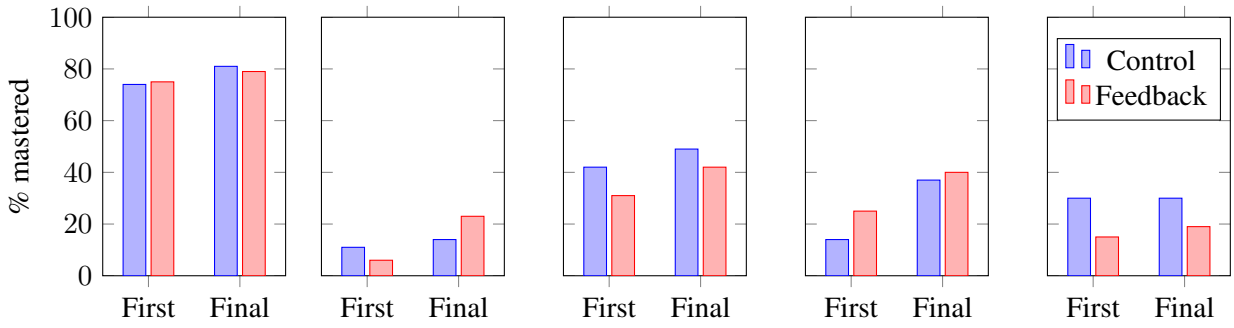
Figure 8: Percentage of students who mastered a criterion according to automatic scoring for each revision step.

Automatic Scoring of E-mail Revisions While the study was initially conducted, only the first and final revision of an e-mail were scored automatically. We later scored each revision automatically according to each criterion in order to check whether improvements indeed mainly occurred after the respective feedback was received. Figure 8 indicates that for the feedback group, this expectation was confirmed, while the control group showed a less pronounced step-wise improvement.

Quality of Automatic Scores To check the automatic scoring performance on the newly collected e-mails, we manually scored the first as well as the final revision of each e-mail after the study was completed. Scoring was performed by a trained annotator who had already been involved in the scoring process of the eRubrix dataset. In doing so, we are able to validate the scoring performance of our automatic scoring model on this new data. For comparison, Table 5 contains cross validation results on the training data in the first line followed by scoring performance for the first and last draft of the e-mails from our intervention studies. For the two criteria *Greeting* and *Interpersonal*, performance is close to the performance in the training data, for *Content* and *Subject* performance deteriorates. For the latter criterion the cause might lie in issues of annotation where a single frequent subject line was scored differently between the texts in the eRubrix dataset and our study. In addition, we found population effects, where the new data contained formulations and lexical elements never encountered during training. The *Style* criterion could only be predicted unreliably in all conditions and was also the criterion with the lowest human-human agree-

Test data	Content		Greeting		Subject		Interpersonal		Style	
	acc	κ	acc	κ	acc	κ	acc	κ	acc	κ
eRubrix - CV	.92	.60	.88	.67	.94	.87	.85	.69	.81	.35
Intervention Study - First Draft	.78	.30	.95	.68	.70	.36	.85	.60	.78	.14
Intervention - Final Draft	.83	.29	.89	.63	.70	.41	.81	.62	.79	.27

Table 5: Scoring accuracy for the language school prompt used in our intervention study. We repeat the cross-validation experiments on the eRubrix data (first line) and then present results for the first and final draft in the intervention study.



(a) Content (b) Greeting (c) Subject Line (d) Interpersonal (e) Register/Style

Figure 9: Improvements for control and feedback group according to the manual scoring.

ment. The two criteria with the best automatic scoring performance (*Greeting* and *Interpersonal*) also showed the highest improvement in the feedback group. We repeated the analyses described in 5.2 for the manual ratings. While the pattern of the results looks similar (see Figure 9), only one out of the five criteria showed statistically significant improvement. The only significant interaction was found for *Greeting and Closing* (χ^2=4.14; p $<$ .05).

6 Discussion & Limitations

One limitation of our automated scoring approach is that for most scoring categories, we feed the whole text into the automatic classification model even though only certain parts are directly relevant (for example, to judge the appropriateness of the closing sentence it would be enough only to consider this particular sentence for scoring). To explore the options for further improvement, we therefore started to collect gold-standard annotations identifying the specific section where each element is located in the text so that we can use a two-stage approach in the future, where we first learn how to segment the text and then classify the appropriateness of the resulting segments.

In our intervention study, we were not able to separate effects of individual feedback components. Therefore we do not know the contribution of evaluative and elaborative feedback components. However, when looking at individual revisions, we saw a clear tendency that students relied on automatic feedback when deciding whether to revise their texts at all. Similarly, we used a very simple binary feedback that could be further improved, e.g. by highlighting relevant parts of an e-mail or by containing more specific hints for improvement.

We also scored only the first and last revision of the email automatically during the intervention study, while feedback (based on the first draft) was provided iteratively for each revision step. It is possible that students improved an aspect of the email that was only addressed later, so that feedback for that criterion was inaccurate at the point in time when the feedback was given. Our post-hoc automatic scores for each revision step (see Figure 8), however, indicate that this was rarely the case. Currently, we also do not know whether improvements will be long-term or whether students will be able to transfer them to new, unfamiliar e-mail writing prompts.

7 Conclusion

We presented a feedback intervention study based on automatic scores for an e-mail writing task

scored according to different criteria. We found that students from the feedback groups improved more than students from the control groups for those two (out of five) criteria where the scoring algorithm worked best. Although much more work into similar directions is needed, especially with respect to the limitations discussed above, our study hints at the general usefulness of automatic scoring in the classroom.

8 Acknowledgements

This work was partially conducted in the framework of CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany, and partially within the KI-Starter project "Explaining AI Predictions of Semantic Relationships" funded by the Ministry of Culture and Science Nordrhein-Westfalen, Germany. We would like to thank our student assistant Sascha Meyer for his contributions to the automatic scoring model.

References

Robert L Bangert-Drowns, Chen-Lin C Kulik, James A Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of educational research*, 61(2):213–238.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898*.

Johanna Fleckenstein, Raja Reble, Jennifer Meyer, Thorben Jansen, Lucas W. Liebenow, Jens Möller, and Olaf Köller. in press. Digitale Schreibförderung im Bildungskontext: Ein systematisches Review. In K. Scheiter and I. Gogolin, editors, *Bildung für eine digitale Zukunft (Edition ZfE, Band XX)*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Steve Graham, Michael Hebert, and Karen R Harris. 2015. Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, 115(4):523–547.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 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*, pages 357–366, Copenhagen, Denmark. Association for Computational Linguistics.

Ken Hyland. 2007. Genre pedagogy: Language, literacy and l2 writing instruction. *Journal of second language writing*, 16(3):148–164.

Stefan D. Keller, Ruth Trüb, Emily Raubach, Jennifer Mayer, Thorben Jansen, and Johanna Fleckenstein. 2023. Designing and validating an assessment rubric for writing emails in English as a foreign language. *Research in Subject-matter Teaching and Learning (RISTAL)*.

Ronald T Kellogg, Alison P Whiteford, and Thomas Quinlan. 2010. Does automated feedback help students learn to write? *Journal of Educational Computing Research*, 42(2):173–196.

Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016. Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72.

Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2010. Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3):391–417.

Danielle S McNamara, Scott A Crossley, Rod D Roscoe, Laura K Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

Edna Holland Mory. 2004. Feedback research revisited. In *Handbook of research on educational communications and technology*, pages 738–776. Routledge.

Susanne Narciss. 2008. Feedback strategies for interactive learning tasks. In *Handbook of research on educational communications and technology*, pages 125–143. Routledge.

Rodney D. Nielsen, Wayne Ward, James Martin, and Martha Palmer. 2008. Annotating students' understanding of science concepts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rodney D Nielsen, Wayne Ward, and James H Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501.

Kimihiro Noguchi, Yulia R Gel, Edgar Brunner, and Frank Konietschke. 2012. nparld: an r software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical software*, 50:1–23.

Corey Palermo and Margareta Maria Thomson. 2018. Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, 54:255–270.

Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Isaac Persing and Vincent Ng. 2016. End-to-End Argumentation Mining in Student Essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394.

D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional science*, 18(2):119–144.

Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Marie Stevenson and Aek Phakiti. 2014. The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, 19:51–65.

Carola Strobl, Emilie Ailhaud, Kalliopi Benetos, Ann Devitt, Otto Kruse, Antje Proske, and Christian Rapp. 2019. Digital support for academic writing: A review of technologies and pedagogies. *Computers & education*, 131:33–48.

David Wade-Stein and Eileen Kintsch. 2004. Summary street: Interactive computer support for writing. *Cognition and instruction*, 22(3):333–362.

Abraham Wald. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Henry George Widdowson. 1978. *Teaching language as communication*. Oxford university press.

Joshua Wilson and Amanda Czik. 2016. Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100:94–109.

Joshua Wilson and Rod D Roscoe. 2020. Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1):87–125.

Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10:3087.

Jinfen Xu and Shanshan Zhang. 2022. Understanding awe feedback and english writing of learners with different proficiency levels in an efl classroom: A sociocultural perspective. *The Asia-Pacific Education Researcher*, 31(4):357–367.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in esol learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43.