Leppänen, Leo; Hellas, Arto; Leinonen, Juho

Piloting Natural Language Generation for Personalized Progress Feedback

Published in:
2022 IEEE Frontiers in Education Conference (FIE)

DOI:
10.1109/FIE56618.2022.9962555

Published: 01/10/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Piloting Natural Language Generation for Personalized Progress Feedback

Leo Leppänen
Department of Computer Science
University of Helsinki
Helsinki, Finland
leo.leppanen@helsinki.fi

Arto Hellas
Department of Computer Science
Aalto University
Espoo, Finland
arto.hellas@aalto.fi

Juho Leinonen
Department of Computer Science
Aalto University
Espoo, Finland
juho.2.leinonen@aalto.fi

***Abstract*—*Full research paper*—We describe the results of a pilot study wherein we applied simple natural language generation methods to produce automated feedback for students of an online course based on student high-level progress data. Experimenting with both personalized and non-personalized feedback, we show that such feedback can be easily produced given access to even rudimentary data regarding student assignment submissions and their correctness. Our results suggest that students perceive automatically generated feedback generally positively and believe it to be useful. Our results also indicate that minor personalization and stylistic alterations in the feedback can have meaningful effects on how the feedback is interacted with and perceived. In particular, we observe that personalized feedback is perceived as being slightly easier to understand and as being better aligned with their progress. Students also felt better about the personalized feedback in comparison to non-personalized feedback. We conclude that the automated generation of personalized textual feedback shows promise as a low-threshold way of increasing student satisfaction. Further research is needed to assess the effect of different types of automated personalized feedback on student performance and behavior.**

***Index Terms*—feedback, automated feedback, personalized feedback, personalization, natural language generation**

I. Introduction

Feedback is a crucial part of learning [1], [2]. It can be given on multiple levels ranging from very specific feedback that targets performance in individual tasks to very generic feedback that broadly considers personal characteristics such as behavior [1], [2]. The level of detail of feedback – or the information content – also influences the effectiveness of feedback [3]. This is, however, not surprising, as the detail can also provide something to act on. As an example, within computing education, there exists plenty of research into automated assessment systems [4], [5] – if such systems would simply state that something was correct (or incorrect), students could struggle more to find what went wrong. Indeed, if the feedback from an automated assessment system targets specific parts of what is currently being worked on or even provides scaffolds [6], [7], students likely need to spend less time on determining the next steps. Naturally, such scaffolds can also be built within assignments themselves [8]–[11].

Computing education researchers have come up with multiple ways to facilitate the process of giving feedback to students. The previously mentioned automatic assessment systems are prime examples of this, but other types of feedback have also been explored. As an example, researchers have studied dashboards and visualizations for providing feedback on process and progress [12]–[15], and created means to more efficiently provide written feedback [16]–[18]. In the former example, the feedback is typically succinct and based on progress, while in the latter example, feedback is built as a part of grading work using rubrics. However, feedback may not always be beneficial as it is, for example, possible that students do not understand it [19]. Indeed, sometimes no feedback at all might be better than say, short textual feedback [12].

In this work, we outline our findings from a study where our objective has been to provide students succinct feedback that targets their current progress, but instead of using a visualization or a dashboard, using written feedback. One novelty in this work is that instead of writing the feedback personally, we have explored the possibility of using lightweight natural language generation for creating the feedback. Acknowledging that the level of feedback and the information content influences the effectiveness of feedback [1], [3], the data used to guide the construction of the feedback has stemmed from a course platform, where the generated feedback has also been provided to the students. In terms of the format of the given feedback, a close match to our study is the rule-based feedback from Gkatzia et al. [20] who, however, considered more aspects of learning, e.g. lecture attendance and students' health.

This article is organized as follows. Next, we outline principles of natural language generation and discuss related works in personalized and adaptive feedback. In Section III, we outline the research methodology, including the research questions and approach as well as a description of the study context, construction of feedback, and collected data. Section IV outlines the results of our study, which are further discussed in Section V. Finally, Section VI provides limitations of this work and Section VII concludes the article and points out future research directions.

II. Background

A. Natural Language Generation

Previous works have found that describing data as text has benefits over other media, such as graphs. For example,

Law et al. [21] found that neonatal Intensive Care Unit nurses tended to "choose more of the appropriate actions when the information was presented as text rather than as graphs." Similarly, textual descriptions have been shown to result in improved decision making under uncertainty [22]. These results indicate that there are, at least in some contexts, concrete benefits to providing information in textual format rather than in alternative formats.

Natural language generation is a well-established area of computer science investigating how various (usually non-linguistic) inputs can be translated into natural language texts using a variety of methods spanning from simple template-filling systems to – more recently – more complex machine learning models [23]. The consensus on the relative benefits of these methods is still developing, but in general it appears that while neural generation methods are able to produce very natural outputs, they also often produce text that is not grounded in the input [24], [25]. They can also be outperformed by hand-engineered systems especially in limited domains [24], [25]. In addition, machine learning methods are dependent on the existence of rather large amounts of training data, which might not be available in all domains, or the production of which might be prohibitively expensive. At the same time, rule-based or template-based methods are viewed as more expensive to create (at least ignoring data engineering costs associated with machine learning methods) and lacking in output variation. Such lack of variation can have negative effects especially in contexts where readers are presented with multiple generated texts in close succession [26].

In the educational setting, natural language generation has been mostly applied in the context of tutoring systems. For example, Di Eugenio et al. [27] investigated various NLG approaches to producing descriptions that teach students to troubleshoot complex systems while Wang et al. [26] produced descriptions of database query execution plans. On the other hand, for example Boyer et al. [28] investigated a tutor system, formulated as a chatbot, that would provide students with programming help.

Another relevant aspect of natural language generation relating to education is personalization, with the hypothesis being that it is better to provide each student (or student archetype) with a personalized textual output than to generate a singular 'one-size-fits-all' output. The practical usefulness of such personalization, however, is not necessarily beneficial. For example, Reiter et al. [29] provided participants both personalized and non-personalized letters with the goal of having the recipients stop smoking. Their personalized letters performed equally well as the non-personalized letters.

B. Personalized and Adaptive Feedback

The influence on feedback on learning is great [1]. However, giving effective feedback can be time-consuming for instructors, and thus ways of automating the process of generating and giving feedback to students have been studied a lot in the context of programming [30]. In programming, students can get feedback in multiple ways – Le [31] proposes a classification of adaptive feedback for programming that comprises of five types of feedback: yes/no feedback, syntax feedback, semantic feedback, layout feedback, and quality feedback. Based on the findings of a systematic literature review into automated programming exercise feedback generation by Keuning et al. [30], most tools focus on identifying student mistakes and fewer on how to fix the problems that have been found. This is unfortunate as quality feedback should indicate the steps that students can take to improve [2].

Not all feedback is effective. For example, compiler error messages can often be confusing to novices who might struggle to understand what they mean [32]; and for feedback to have an effect, it should be understandable for students [2]. Thus, in the context of programming, prior work has examined ways of giving automated *actionable* feedback. One focus of this work has been automatic feedback in the form of automatically generated hints [33]–[36]. Rivers et al. [33]–[35] propose a solution space based approach where the current state of a student's program is compared to program states in paths that lead to correct solutions where the hint suggests a change that guides the student towards a correct solution. A similar approach was taken by Keuning et al. [36] who infer the strategy students are taking to solve the problem and guide the student towards a model solution that corresponds to their chosen strategy.

As proposed by Ott et al. [2], "Ideally, feedback should take the learner's characteristics and abilities into account". A similar suggestion to personalize feedback was recently made by Wang et al. [37] who studied Step Tutor, which gives students' example-based feedback on programming tasks. Prior work into personalizing feedback has found promising results. For example, Marwan et al. [38] studied autograder based adaptive immediate feedback presented within a block-based programming environment. The feedback in the system was personalized based on student actions such as whether they had been inactive or not. Their results suggested that high school students who got the adaptive personalized immediate feedback had increased intentions to persist in CS as well as greater engagement and increased learning. Similarly, Kochmar et al. [39] found increased learning gains for students who had personalized feedback that was generated with natural language processing methods. Voghoei et al. [40] sent personalized feedback emails that had forecasts about students' grades and found that especially students in the middle of the grade-range improved their performance as a result; and additionally found that students reported decreased stress as a result of the personalized feedback emails.

III. Methodology

A. Context

The study was conducted in a Web Software Development course offered by Aalto University in Finland. The Web Software Development course is a 5 ECTS course, corresponding to approximately 135 hours of study. Students attending the course are mainly second-to-third year Bachelor's level students with Computer Science either as a Major subject or

as a Minor subject, although the course is taken also by many Master's level students.

In Fall 2021, when this study was conducted, in part due to the COVID-19 situation, the course was given as a continuously available online course, where students could take the course in a way that would best fit their schedule, and students could continue in the course even if they would have to take a longer absence. Participants in the course were supported through an online discussion platform as well as through an online help request functionality that is embedded in the course platform.

The course follows an online textbook format with interleaved assignments and quizzes. The assignments and quizzes are automatically assessed. In addition, students have two larger projects, which are self-, peer-, and instructor-reviewed. The grading of the course is based on the number of completed assignments and quizzes; the projects are graded pass or fail, where passing a project means fulfilling all project requirements. Students can choose to complete the course with either one or two projects, where completing the course with only one project limits the grading so that the highest grades are not available.

Students receive feedback from automated assessment, which essentially uses test suites to check the correctness of the programs, as well as from project reviews. In addition, in the present study, we looked into automatically generated textual feedback.

B. Randomized controlled trial

We created a randomized controlled trial, where students were divided into three groups. The first group saw no feedback, the second group saw automatically generated generic feedback, and the third group saw automatically generated personalized feedback. For both the second and the third group, the feedback was always constructed based on students' behavior in the previous chapter of the course materials.

Assignment into a group was performed automatically when a student entered a page with feedback; a student was always assigned to the group with the least students (randomly selected out of the groups with the least students in the case where no single group had the least students). Through this, our objective was to have evenly distributed groups with somewhat even progress between the groups.

C. Feedback generation

Based on our observations regarding the properties of various approaches to natural language generation (see Section II) we determined that a template-based approach provided a suitable starting point given the domain. The use of e.g. neural generation models would have necessitated the production of large training corpora – which did not exist for the feedback generation task – and the tendency to produce output not grounded in input (i.e. producing feedback that does not match the student) would be a significant problem, while the increased textual variety would not have been a great benefit.

In the experiment, feedback was constructed based on five factors each describing student behavior in the previous chapter:

- Percentage of points from previous chapter.
- Completed assignments and average attempts per assignment.
- Submissions to assignments after having already completed them.
- Unfinished assignments.
- Skipped assignments.

A feedback phrase was constructed from each of the five factors. For the generic feedback, the feedback phrase was always the same, while for the personalized feedback, the feedback phrase depended on the factor value. The feedback also might provide suggestions on study, depending on the item. To provide concrete examples, the two following quotes describe generic feedback that one might see for the percentage of points from previous chapter and the completed assignments and average attempts per assignment.

> You've gained 72% of the points available in the last part. Always aim to complete all the assignments for maximal learning!
>
> In the previous part, you completed 8 assignments. On average, you have 1.5 incorrect attempts for every correct attempt. Testing out solutions on your own computer before submitting them is a good strategy!

The following quotes show three different personalized feedback texts for the percentage of points from previous chapter, with the system selecting which item to show depending on the current percentage of points that the student had.

> So far, you've gained less than 70% of the total points available in the previous part. We recommend that you return to the previous part before continuing to this part.
>
> You've gained 72% of the total points available in the previous part. Good work! We recommend that you return to the previous part and finish the remaining assignments at some point as practice is an important part of learning.
>
> You've gained 88% of the total points available in the previous part. Great work! We recommend that you return to the previous part and finish the remaining assignments before finishing the course, as practice is an important part of learning.

The exact boundaries that were used for creating the personalized feedback were tuned by the course instructor, who has been responsible for the course for the past two years. The key logic and rules used to generate the feedback is provided as an online supplement[1].

D. Data

The feedback was made available to the students at the top of the first page of the chapter following the one that the

[1]https://osf.io/82rbz/?view_only=c1f692336d874b11a64e170c19af429c

feedback was constructed on. At the top of the page, students saw a button with a text that asked them to press the button to see feedback on their progress in the previous part. After pressing the button, students were shown the textual feedback, and were also given a form that asked for counter-feedback on the generated feedback. Students were not informed that the feedback was automatically generated.

The button for viewing the feedback was different for the two groups, with the generic treatment group being shown the prompt "Hi! We've got some feedback for you. Click here to show it!" while the personalized treatment group was shown the prompt "Click here to view feedback for part (previous part identifier)." We will discuss some implications of this in Section V.

The counter-feedback contained the following items, each answered using a Likert-like scale from *1 = Strongly disagree* to *5 = Strongly agree*.

- This feedback was useful.
- I understood the feedback.
- I think that the feedback matched my progress well.
- I feel good about the feedback.

For the present study, we collected data on (1) students' pressing the button that lead to feedback being shown, (2) counter-feedback from students, and (3) students course points and submissions. The data was collected between November 2021 and January 2022.

E. Research questions and approach

Using the data from the randomized controlled trial, we answer the following research questions.

RQ1. How do students perceive personalized and non-personalized feedback?

RQ2. How does the personalized and non-personalized feedback affect student behavior?

For RQ1, we compare the counter-feedback given by students with the generic feedback to the counter-feedback given by students with the personalized feedback. As the data is ordinal (Likert), we use a Mann-Whitney U test to compare groups. In addition to reporting the Mann-Whitney U statistic and p-value, we report effect sizes using Cohen's d.

For RQ2, we compare the three groups (no feedback, generic feedback, personalized feedback) regarading the number of total exercise points they received in the course on average. Additionally, we use a chi-squared test to examine whether there were differences between the two groups that received feedback (generic and personalized) in the number of times they gave counter-feedback.

For both RQs, we only analyze students who had received at least a single point from exercises as the feedback was related to the number of received exercise points.

IV. Results

A. Descriptive statistics

Table I shows descriptive statistics related to the experiment. In total, there were 449 students who had at least a single point from the course assignments and were thus included in the analysis. Out of these, a total of 132 viewed the feedback at least once. The group assigned into the personalized treatment viewed the feedback on average 3.3 times ($\sigma = 4.1$, median 2), while the generic treatment group viewed the feedback on average 3.8 times ($\sigma = 3.5$, median 3). These distributions are not statistically significantly different per a Kruskal-Wallis test (H=2.25, p=0.13). Not all of those who viewed the feedback answered the associated counter-feedback questionnaire. In the personalized treatment group, 37 participants (65%) filled the questionnaire for at least one feedback text, while the relevant number for the generic treatment group was 57 participants (76%). On average, the personalized treatment group provided feedback on 2.11 feedback texts, whereas those in the generic feedback group filled the questionnaire for 3.05 feedback texts.

B. Student perceptions

Figures 1 and 2 show the distributions of students' answers to the counter-feedback questionnaire questions separately for the group that received generic feedback (Figure 1) and the group that received personalized feedback (Figure 2). From the figures, one can see that overall, for both groups, students generally answered more positively than negatively. Based on visual inspection of the figures, the group with the personalized feedback seems to agree and strongly agree more on the counter-feedback statements, while the generic feedback group has more "neither agree nor disagree" answers.

To complement the visual analysis, we examined differences between the groups that had feedback using Mann-Whitney U tests for statistical significance. To combat the multiple comparisons problem, we correct the p-values with a Bonferroni-correction. Since there are four counter-feedback questions, we use a significance threshold of a = 0.05/4 = 0.0125.

Looking at students' general feelings about the feedback ("I feel good about the feedback"), there was a statistically significant difference between the groups (U = 5362.5, p = 0.002). The median for both groups was "Agree". Cohen's d effect size was 0.37, with those in the personalized group giving more positive answers on average.

For the statement "I think that the feedback matched my progress well", the Mann-Whitney U test also signaled a statistically significant difference (U = 5024, p = 0.0001). The median for both groups was again "Agree". Cohen's d effect size was 0.45 revealing that the personalized group again gave more positive answers on average.

Regarding understanding of the feedback ("I understood the feedback"), there was again a statistically significant difference between the groups (U = 4185, p = 1.05e-07). The median for the generic group was "Agree" while it was "Strongly agree" for the personalized group. Cohen's d effect size between the groups was 0.62 with the personalized group agreeing more on average with the statement.

Lastly, for the statement related to the usefulness of the feedback ("This feedback was useful"), the difference between the groups was not statistically significant (U = 6020.5, p = 0.06). The median for both groups was "Agree". Cohen's d

TABLE I
DESCRIPTIVE STATISTICS.

	Total	No feedback	Generic feedback	Personalized feedback
Participants	449	142	154	153
Opened feedback	132	-	75	57
Feedback texts viewed (mean / std / median)	3.6 / 3.7 / 2	-	3.8 / 3.5 / 3	3.3 / 4.1 / 2
Gave counter-feedback at least once	94	-	57	37
Total counter-feedback given	252	-	174	78
Counter-feedback given per student (mean / std / median)	2.7 / 3.0 / 1	-	3.1 / 3.3 / 2	2.1 / 2.4 / 1

effect size was 0.17, again with the personalized group being in greater agreement with the statement.

C. Effects on student behavior

We also considered students' behavior. Firstly, we consider progress as measured by course points. The average points and point standard deviations (marked with σ) of all those students who gained at least one point and who were assigned to one of the three groups were as follows.

- No feedback: 2257.8 (σ=1785.1)
- Generic feedback: 2325.7 (σ=1834.9)
- Personalized feedback: 2392.7 (σ=1788.0)

At the time of the writing of this article, the maximum possible points for the course is 4770. Due to the continuous nature of the course, i.e. students could have been in different parts of the course when the intervention began and students could also continue after the intervention ended, any inferences made based on the points should be taken with a grain of salt. Nevertheless, a Kruskal-Wallis test indicates the groups are not statistically significantly different ($p = 0.78$).

Secondly, regarding students' behavior, we consider whether the type of feedback had an effect on the number of times students gave counter-feedback. As mentioned in the beginning of the results section, on average, the personalized treatment group provided counter-feedback 2.11 times, whereas those in the generic feedback group filled the counter-feedback questionnaire 3.05 times. The number of participants filling the questionnaire, controlled by the number of those who viewed the feedback, is statistically significantly different between the two treatment groups (p=0.023) using a χ^2 test with Yates' correction for continuity, meaning that the generic feedback elicited statistically significantly more counter-feedback from the participants.

V. DISCUSSION

We interpret our results, as described above, as providing three primary insights into automated generation of high-level feedback from student process data. First, we observe that personalized student feedback can be generated with relatively simple natural language generation methods. Our generation is based on a very simple template-filling approach. As such, it appears to us that these types of adaptive systems could be easily integrated into almost any course with information on student progress (submissions and their correctness) as there is no need for e.g. the voluminous training data associated with recent neural natural language generation models. As the texts are interacted with relatively infrequently, we believe that these types of lightweight template-based approaches are more than sufficient, and that the observations of Wang et al. [26] relating to boredom stemming from frequent viewing of highly similar texts do not apply to this context.

Second, our results indicate that students perceive automated feedback positively. Taken together with previous works indicating that student feedback is associated with various positive effects [1], we interpret this as a positive signal that the automated generation of feedback texts would very likely result in improved student performance over a baseline situation where a time-pressed teacher is unable to provide the students with any feedback.

Third, the results suggest that increased personalization of the automated feedback results in better comprehension of the feedback items, an increased perception of the feedback matching the recipient's progress, as well as a generally more positive perception of the feedback. These results are in line with Gkatzia et al. [20] who also found that a rule-based personalized feedback system was preferred by students over other types of systems. Given that these results were obtained already from an initial pilot study – i.e. the style and content of the personalized feedback was not iterated upon – we interpret this as indicating that further study should be directed towards identifying what are the most salient features and stylistic details of the feedback.

We found that students with the generic feedback gave more counter-feedback compared to those with personalized feedback. We hypothesize that the difference is attributable to the relatively minor stylistic difference between the invitations to interact with the feedback. This would be well-aligned with previously published results relating to the *emotional design effect* [41], and is one more aspect of the feedback that warrants further study and iteration.

Previous works related to natural language generation in a news text context indicate that the attribution of text authorship is complicated. In one study, respondents were divided in attributing the authorship of an automatically generated news text to the computer program that created it, to the team of humans who created the system, and a significant proportion being fundamentally unable to attribute authorship at all [42]. In our view, this raises two important questions. First, whether students attribute the automatically generated feedback texts to the teacher of the course, and second, how this affects both the student-teacher relation and the student-feedback relation. On one hand, it is possible that some students might be more likely

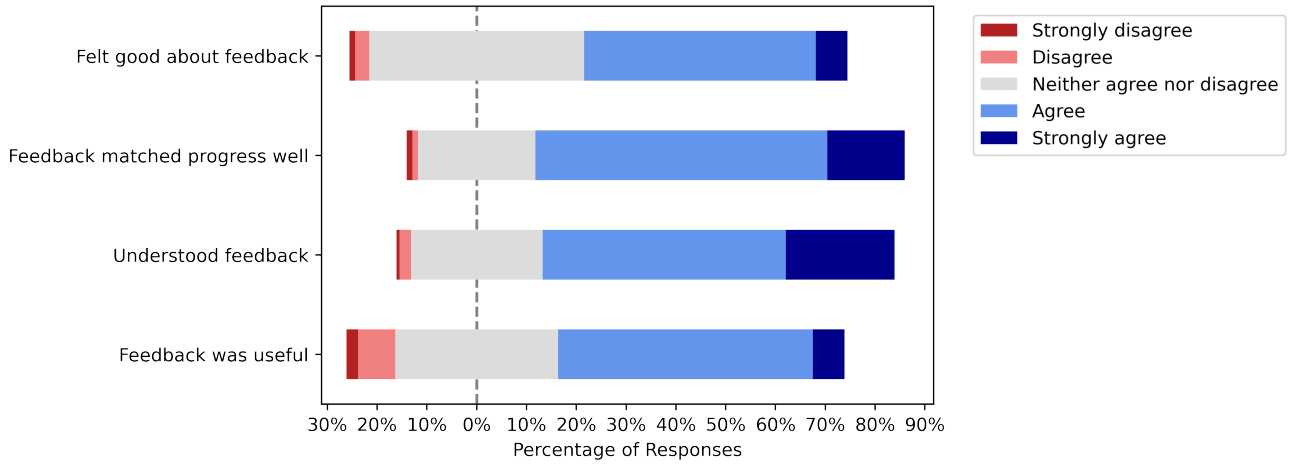

Fig. 1. Distribution of responses to the feedback questions for the generic feedback group.

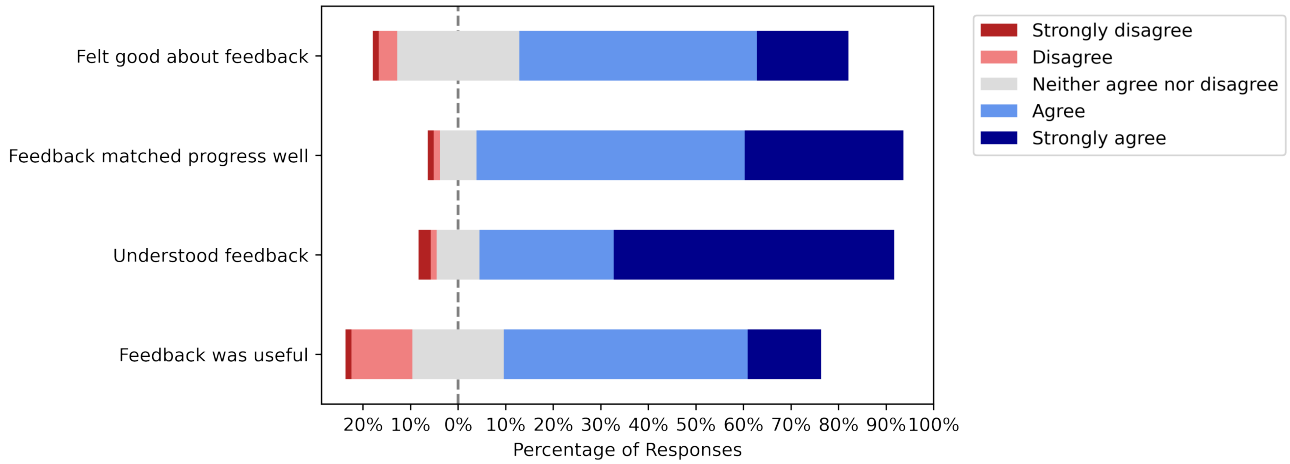

Fig. 2. Distribution of responses to the feedback questions for the personalized feedback group.

to discount the automatically generated feedback items, while on the other hand a potentially increased separation between the teacher and the feedback might result in the students viewing the feedback as more objective and more 'safe', as any possible negative feedback is less associated with a real person.

VI. Limitations

It is notable that the pilot study described above is just that – a pilot study – and comes with several important limitations. First, the results reflect a first iteration of a very lightweight system, meaning that significantly improved results would likely be obtainable if the system was further iterated upon.

Second, the system was introduced into a continuously running online course, which means that the participants were in different parts of the course when the trial began. For example, some might have joined the trial extremely late in the course, which would cause a ceiling effect into the number of times they were able to interact with the system, as well as in terms of the points they would be able to obtain before finishing the course. Similarly, some participants would have joined the experiment very close to it's end, again resulting in a similar ceiling effect.

Furthermore, we acknowledge that the results reported in Section IV include means and standard deviations of ordinal Likert scale data. We provide this information to give the reader additional insight into the data, but caution that care should be taken when interpreting these values due to the ordinal nature of the underlying data. On the other hand, the statistical tests conducted, e.g. Mann-Whitney's U tests and Kruskal-Wallis' H-tests, are suitable for ordinal data. Finally, we note the Bonferroni corrections applied in Section 4. First, we applied these corrections within families of related tests, rather than globally. Second, the correction is commonly considered as being potentially overly conservative in cases where the test statistics are positively correlated within families. These two balancing factors should be taken into consideration when interpreting the statistical significance of the results.

VII. Conclusion

In this paper we have described a pilot study on automated generation of textual feedback for students based on student progress data. Our experiences indicate that textual feedback can be generated with very lightweight natural language generation solutions, that students perceive feedback generally positively, and that personalizing the feedback results in more understandable feedback which is also viewed more positively.

Due to the limited nature of this pilot study, further work is needed to investigate what style and content are best suited for such automated feedback systems and how the feedback affects student performance both in the long and short term. Finally, our experiment raises additional questions regarding how the students perceive the authorship of the automatically generated feedback and how the provision of automatically generated feedback affects student-teacher relations. We intend to investigate these aspects in future work.

References

[1] J. Hattie and H. Timperley, "The power of feedback," *Review of educational research*, vol. 77, no. 1, pp. 81–112, 2007.

[2] C. Ott, A. Robins, and K. Shephard, "Translating principles of effective feedback for students into the cs1 context," *ACM Transactions on Computing Education (TOCE)*, vol. 16, no. 1, pp. 1–27, 2016.

[3] B. Wisniewski, K. Zierer, and J. Hattie, "The power of feedback revisited: a meta-analysis of educational feedback research," *Frontiers in Psychology*, vol. 10, p. 3087, 2020.

[4] P. Ihantola, T. Ahoniemi, V. Karavirta, and O. Seppälä, "Review of recent systems for automatic assessment of programming assignments," in *Proceedings of the 10th Koli calling international conference on computing education research*, 2010, pp. 86–93.

[5] K. M. Ala-Mutka, "A survey of automated assessment approaches for programming assignments," *Computer science education*, vol. 15, no. 2, pp. 83–102, 2005.

[6] A. Vihavainen, T. Vikberg, M. Luukkainen, and M. Pärtel, "Scaffolding students' learning using test my code," in *Proceedings of the 18th ACM conference on Innovation and technology in computer science education*, 2013, pp. 117–122.

[7] P. E. Anderson, T. Nash, and R. McCauley, "Facilitating programming success in data science courses through gamified scaffolding and learn2mine," in *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, 2015, pp. 99–104.

[8] S. Marwan, P. Shabrina, A. Milliken, I. Menezes, V. Catete, T. W. Price, and T. Barnes, "Promoting students' progress-monitoring behavior during block-based programming," in *21st Koli Calling International Conference on Computing Education Research*, 2021, pp. 1–10.

[9] L. E. Margulieux and R. Catrambone, "Improving problem solving with subgoal labels in expository text and worked examples," *Learning and Instruction*, vol. 42, pp. 58–71, 2016.

[10] L. E. Margulieux, M. Guzdial, and R. Catrambone, "Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications," in *Proceedings of the ninth annual international conference on International computing education research*, 2012, pp. 71–78.

[11] J. Joentausta and A. Hellas, "Subgoal labeled worked examples in k-3 education," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018, pp. 616–621.

[12] K. Ilves, J. Leinonen, and A. Hellas, "Supporting self-regulated learning with visualizations in online learning environments," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018, pp. 257–262.

[13] H. He, Q. Zheng, and B. Dong, "Learnerexp: exploring and explaining the time management of online learning activity," in *The World Wide Web Conference*, 2019, pp. 3521–3525.

[14] S. Govaerts, K. Verbert, E. Duval, and A. Pardo, "The student activity meter for awareness and self-reflection," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 2012, pp. 869–884.

[15] T. Auvinen, L. Hakulinen, and L. Malmi, "Increasing students' awareness of their behavior in online learning environments with visualizations and achievement badges," *IEEE Transactions on Learning Technologies*, vol. 8, no. 3, pp. 261–273, 2015.

[16] T. Ahoniemi, E. Lahtinen, and T. Reinikainen, "Improving pedagogical feedback and objective grading," in *Proceedings of the 39th SIGCSE technical symposium on Computer science education*, 2008, pp. 72–76.

[17] L. Gillam, D. Bush, G. Qin, and N. Newbold, "Automating feedback: The cafex2 project," in *10 th Annual Conference of the Subject Centre for Information and Computer Sciences*, 2009, p. 70.

[18] T. Auvinen, "Rubyric," in *Proceedings of the 11th Koli Calling International Conference on Computing Education Research*, 2011, pp. 102–106.

[19] C. Glover and E. Brown, "Written feedback for students: too much, too detailed or too incomprehensible to be effective?" *Bioscience education*, vol. 7, no. 1, pp. 1–16, 2006.

[20] D. Gkatzia, H. Hastie, S. Janarthanam, and O. Lemon, "Generating student feedback from time-series data using reinforcement learning," in *Proceedings of the 14th European Workshop on Natural Language Generation*, 2013, pp. 115–124.

[21] A. S. Law, Y. Freer, J. Hunter, R. H. Logie, N. McIntosh, and J. Quinn, "A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit," *Journal of clinical monitoring and computing*, vol. 19, no. 3, pp. 183–194, 2005.

[22] D. Gkatzia, O. Lemon, and V. Rieser, "Data-to-text generation improves decision-making under uncertainty," *IEEE Computational Intelligence Magazine*, vol. 12, no. 3, pp. 10–17, 2017.

[23] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[24] S. Wiseman, S. M. Shieber, and A. M. Rush, "Challenges in data-to-document generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2253–2263.

[25] O. Dušek, J. Novikova, and V. Rieser, "Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge," *Computer Speech & Language*, vol. 59, pp. 123–156, 2020.

[26] W. Wang, S. S. Bhowmick, H. Li, S. Joty, S. Liu, and P. Chen, "Towards enhancing database education: Natural language generation meets query execution plans," in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 1933–1945.

[27] B. Di Eugenio, D. Fossati, D. Yu, S. Haller, and M. Glass, "Aggregation improves learning: experiments in natural language generation for intelligent tutoring systems," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 2005, pp. 50–57.

[28] K. E. Boyer, R. Phillips, A. Ingram, E. Y. Ha, M. Wallis, M. Vouk, and J. Lester, "Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach," *International Journal of Artificial Intelligence in Education*, vol. 21, no. 1-2, pp. 65–81, 2011.

[29] E. Reiter, R. Robertson, and L. M. Osman, "Lessons from a failure: Generating tailored smoking cessation letters," *Artificial Intelligence*, vol. 144, no. 1-2, pp. 41–58, 2003.

[30] H. Keuning, J. Jeuring, and B. Heeren, "A systematic literature review of automated feedback generation for programming exercises," *ACM Transactions on Computing Education (TOCE)*, vol. 19, no. 1, pp. 1–43, 2018.

[31] N.-T. Le, "A classification of adaptive feedback in educational systems for programming," *Systems*, vol. 4, no. 2, p. 22, 2016.

[32] B. A. Becker, P. Denny, R. Pettit, D. Bouchard, D. J. Bouvier, B. Harrington, A. Kamil, A. Karkare, C. McDonald, P.-M. Osera *et al.*, "Compiler error messages considered unhelpful: The landscape of text-based programming error message research," in *Proceedings of the working group reports on innovation and technology in computer science education*, 2019, pp. 177–210.

[33] K. Rivers and K. R. Koedinger, "Automatic generation of programming feedback: A data-driven approach," in *The First Workshop on AI-supported Education for Computer Science (AIEDCS 2013)*, vol. 50, 2013.

[34] ——, "Automating hint generation with solution space path construction," in *International Conference on Intelligent Tutoring Systems*. Springer, 2014, pp. 329–339.

[35] ——, "Data-driven hint generation in vast solution spaces: a self-improving python programming tutor," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 1, pp. 37–64, 2017.

[36] H. Keuning, B. Heeren, and J. Jeuring, "Strategy-based feedback in a programming tutor," in *Proceedings of the Computer Science Education Research Conference*, 2014, pp. 43–54.

[37] W. Wang, Y. Rao, R. Zhi, S. Marwan, G. Gao, and T. W. Price, "Step tutor: Supporting students through step-by-step example-based feedback," in *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*, 2020, pp. 391–397.

[38] S. Marwan, G. Gao, S. Fisk, T. W. Price, and T. Barnes, "Adaptive immediate feedback can improve novice programming engagement and intention to persist in computer science," in *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 2020, pp. 194–203.

[39] E. Kochmar, D. Do Vu, R. Belfer, V. Gupta, I. V. Serban, and J. Pineau, "Automated personalized feedback improves learning gains in an intelligent tutoring system," in *International Conference on Artificial Intelligence in Education*. Springer, 2020, pp. 140–146.

[40] S. Voghoei, N. H. Tonekaboni, D. Yazdansepas, S. Soleymani, A. Farahani, and H. R. Arabnia, "Personalized feedback emails: A case study on online introductory computer science courses," in *Proceedings of the 2020 ACM Southeast Conference*, 2020, pp. 18–25.

[41] R. E. Mayer and G. Estrella, "Benefits of emotional design in multimedia instruction," *Learning and Instruction*, vol. 33, pp. 12–18, 2014.

[42] L. Henrickson, "Natural language generation: Negotiating text production in our digital humanity," in *Proceedings of the Digital Humanities Congress 2018*, 2019.