9th International Conference on Higher Education Advances (HEAd'23)
Universitat Politècnica de València, València, 2023
DOI: http://dx.doi.org/10.4995/HEAd23.2023.16195

Adverse effects of personalized automated feedback

J. Riezebos, N. Renting, R. van Ooijen, A.J. van der Vaart
University of Groningen, The Netherlands.

Abstract

In large classes with hundreds of students, it is rarely feasible to provide students with individual feedback on their performance. Automatically generated personalized feedback on students' performance might help to overcome this issue, but available empirical effect studies are inconclusive due to lack of methodological rigor. This study uses a repetitive randomized control experiment to explore whether automatically generated feedback is effective and for which students. Our results indicate that feedback does not have a positive effect on performance for all students. Some groups benefit from receiving personalized feedback, while others do not perform better than the control group. Students that perform average benefit most from receiving personalized feedback. However, lower-scoring students who received feedback tend to have lower attrition rates and if they participate at the final exam, their performance is not higher than the control group. Therefore, providing automated feedback is not something that should be undertaken mindlessly.

Keywords: *Automated feedback; summative assessment; randomized control experiment; empirical study; adverse effects.*

1. Introduction

Feedback can be one of the most powerful learning tools in education, but it is difficult to use in some higher educational settings (Hattie & Timperley, 2007). Especially in large undergraduate classes with hundreds of students, it is rarely feasible to provide students halfway through the module with individual feedback on their performance and enable them to act upon it when preparing for the final exam. Technological innovation holds the promise to help overcome this issue, as it enables automatically generated personalized feedback on students' performance. Previous studies that have explored the effects of automatically generated feedback on student performance have shown mixed results and often focused only on smaller learning tasks. Moreover, they lack randomized controlled designs (Morris et al., 2021). We provide a rigorous study that explores in greater detail whether automatically generated feedback halfway through the module is effective and for which kind of student.

A commonly deployed technique to provide students with insight on their mastery of the subject is a midterm some weeks before the final exam. This midterm can be formative or low stake summative and poses an important opportunity for personalized feedback, as it generates data on how each student performs on the different learning objectives, while there is still time to improve before the end of the module. We experiment with a feedback system that automatically generates personalized feedback for students on each of the learning objectives of the module that are tested in both the midterm and final exam. The feedback concerns students' current performance, what is expected from them, and how they should proceed from here on. Feedback is formulated for each learning objective at three levels, for students scoring below standard, standard or above standard. Shortly after the midterms, students receive an automatically generated e-mail with their grade and personalized feedback. However, they do not know that the email is automatically generated, as it is send from the lecturer's account and formulated as a personal message.

We tested this feedback system in three large (n>300 students) first year bachelor modules in business administration. These were all technical modules (Management Science (MS), Supply Chain Operations (SCO), and Statistics (Stat)) that require deep learning strategies as students really need to understand and apply the material rather than recall information. For each module, students were randomly assigned to either the experimental group, receiving an email with their grade and personalized feedback, or the control group, receiving an email with only their grade. Our main outcome measures are attrition rate and performance on the final exam (i.e., not the final grade, as that includes the midterm grade as well). We present effects of receiving automated feedback for low, average and high performing students, for which we used the 33% tertiles as boundaries.

2. Experimental setup

For the field study, we selected three first year modules of a single programme of study. The modules were similar in their assessment plan, as they all included a final exam and a midterm, i.e. an intermediate low stake exam 3-4 weeks before the final exam. All learning objectives tested during the intermediate exam were retested during the final exam. The purpose of the intermediate exams was therefore aimed at informing students on their mastery of the subject and the type of exam they could expect for this module. The midterm is a low stake summative test, as the weigth of the midterm is much lower than the weight of the final exam. This allows students to treat it as a kind of formative assessment, since they could make up a bad achievement on the midterm during the final exam. At the start of the year, all 318 freshmen were asked for consent in participating in an experimental study and to give permission in using their outcomes. The students were not informed about the details and purpose of the study to minimize the risk that the outcomes of the experiment are influenced, because the students are aware that they take part in an experiment (Levitt & List, 2009) and could learn from other's individual feedback. In total 88% of the students gave consent to participate in the study. Upon participation in a midterm, we randomly assigned them to either the treatment or control group. Students who did not gave consent were left out of the study. Note that not all students who participated in the midterm decided to participate in the exam. Hence, the attrition rates of treatment and control group are of interest, as this decision may have been affected by the treatment. After all three midterms had been offered, a within subjects approach with repeated measurements has been applied to draw conclusions on the effectiveness of the treatment, notwithstanding the differences between the repeated measurements, as the modules are different.

2.1. Personalized automated feedback

The intermediate exams consisted of open (short answer) and/or closed questions (multiple choice). For each graded element of the midterm (i.e., question or sub-question), we asked the lecturers to identify the learning objective that was tested by that (sub)question. Per module, 6-10 learning objectives were tested. As an intermediate exam could consist of 20-40 (sub)questions, each learning objective was measured using on average almost 4 questions.

Students who would not receive feedback for this midterm (due their assignment to the control group or their lack of consent) did receive a personal email with their name and the grade on their exam and some general information (e.g., time left untill the final exam). Students in the treatment group were sent a personal email at the same moment as the other students, but in addition to their grade they received information that we denote as personalized feedback. Feedback started postive by denoting all learning objectives that were well-mastered (above standard). Next, the learning objectives that scored standard and might need some more attention were addressed including suggestions per learning objective, and finally the learning objectives that were below standard and for which substantive improvements were required were addressed, including more extensive suggestions what to do to achieve these improvements.

3. Results

3.1 Descriptive statistics

All students enrolled for the modules were invited to participate, although for this study we were only interested in freshmen students. Their was no random selection from the population involved, but only students that participated in a midterm and the final of a module could be measured. Hence, per module we registered whether the student participated in a midterm and could randomly be assigned to the treatment or control group based on the consent they had provided. For module 2a Supply Chain operations (SCO) and 2b Statistics (Stat), only freshmen who were assigned to a control or treatment group in module 1 Management Science (MS) have been considered, as we use a repeated measurement study design.

Table 1. Partipant flow.

Module	Total population (#freshmen with consent)	Midterm participants (#freshmen with consent)	Treatment group (#final exam)	Control group (#final exam)	No consent or midterm
1. MS	342 (290)	335 (254)	143 (134)	146 (141)	53
2a. SCO	305 (255)	286	128 (118)	158 (147)	19
2b. Stat	319 (231)	270	128 (118)	142 (137)	49

3.2 Attrition effects

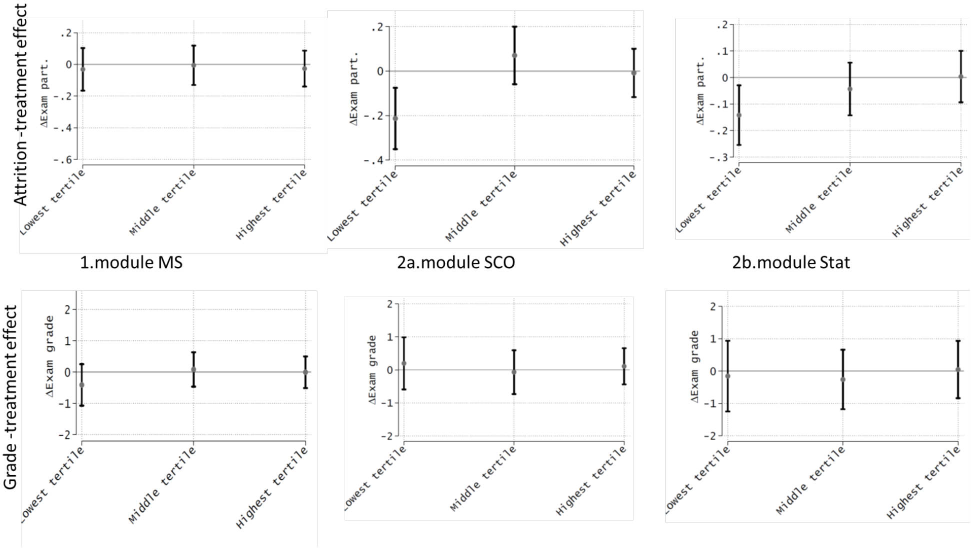
Table 1 shows percentage of freshmen with consent that participated in the midterm (i.e., assigned to either the treatment or control group) and attrited for the final exam. The attrition differences that we found show a much larger attrition in the treatment group. Further examination revelead that these differences cannot be attributed to confounding variables, such as age, gender, high school GPA, et cetera. However, for students that scored low on the midterm (lowest 33%) AND received feedback we found a significant effect in the last

Table 2. Attrition effect of feedback after midterm for each module.

Module	Attrition in control group	Attrition in treatment group
1. MS	3.7%	6.1%
2a. SCO	5.6%	8.9%
2b. Stat	3.7%	8.1%

two modules. Students with a similar low score who did not receive feedback had a higher partcipation in the exam. Hence, feedback to lower scoring students has a significant attrition effect, resulting in higher attrition as a result of personalized specific feedback.

The results are graphically depicted in the upper part of Figure 1, which shows dfferences with the control group. If the boxplot crosses the 0-line, there is no significance treatment effect at 0.05 level.



1.module MS

2a.module SCO

2b.module Stat

Figure 1 Treatment effects on Attrition level (upper part) and Grade of final exam (lower part).

3.3 Grade performance effects

We did not find significant effects of providing personalized feedback (treatment) on the final exam grades in a module. The average grade received by students in the control group was similar to the grade of students in the treatment group. We had expected that the group of students who could benefit most from feedback, i.e. who received suggestions for improvement for only a few learning objectives and hence scored average in the midterm, would show a significant effect of the feedback received. But our data did not confirm this hypothesis, for neither of the modules. Moreover, we did not found a significant performance difference between the lower scoring treatment group students and their control-group counterparts, notwithstanding the higher attrition rates in the treatment group. We denote this as an adverse effect of feedback, as personalized feedback has had no clear impact on performance, only on attrition rate.

4. Discussion and Conclusions

The randomized controlled experiment that we performed leads us to two conclusions. First, we have not been able to demonstrate a positive or negative effect of personalized feedback on the grade performance of the first year students. Next, we have been able to identify an adverse effect of personalized feedback in terms of attrition level in two of the three modules, both located after the students had already taken the first module in their study program.

These conclusions ask for discussion and reflection, as theory suggests mainly positive effects of personalized feedback. Has the experiment correctly been designed? We believe it is, as we followed a similar setup as has been used in other repeated measurements field studies (Levitt & List, 2009). Has the treatment, i.e. the formulation of the feedback, been provided correctly? We followed guidelines of Hattie & Timperley (2007) and Núñez-Peña et al. (2015), but we made the choice to provide feedback based on the achievement of module learning objectives, not on the detailed level of the question itself. Or did we identify an adverse effect of feedback that should lead us to reconsider the power of feedback?

References

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, *53*(1), 1–18. https://doi.org/10.1016/j.euroecorev.2008.12.001

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), 1–26. https://doi.org/10.1002/rev3.3292

Núñez-Peña, M. I., Bono, R., & Suárez-Pellicioni, M. (2015). Feedback on students' performance: A possible way of reducing the negative effect of math anxiety in higher education. *International Journal of Educational Research*, *70*, 80–87. https://doi.org/10.1016/j.ijer.2015.02.005