2018 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
STEAM - Science, Technology & Engineering, Arts, Mathematics & Education
June 6 - 8, 2018 Prince Waikiki, Honolulu, Hawaii

Assessing the Effectiveness of Automated Personalized Feedback in an Undergraduate Biology Course

Schmidt, Mathew et al
University of Saskatchewan
Saskatoon, Saskatchewan
Canada

Mr. Matthew Schmidt
Educational Psychology and Special Education Department
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
Dr. Amin Mousavi
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
Dr. Vicki Squires
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
Dr. Kenneth Wilson
University of Saskatchewan
Saskatoon, Saskatchewan
Canada

Assessing the Effectiveness of Automated Personalized Feedback in an Undergraduate Biology Course

Synopsis:

This paper compares the effectiveness of the implementation of two automated feedback systems in an undergraduate biology course at the University of Saskatchewan. Students were treated with either generic feedback, or personalized feedback. Personalized feedback took a learning analytic approach, and was tailored to individual students based on 40 distinct attributes. The effectiveness of these feedback interventions and their future potential are discussed.

Assessing the Effectiveness of Automated Personalized Feedback in an Undergraduate Biology Course

Matthew Schmidt, Amin Mousavi, Vicki Squires, and Kenneth Wilson

University of Saskatchewan

Introduction

In the context of education, the term formative feedback refers to a process where information is provided to a learner with the intention of modifying the learner's behavior in a beneficial way. More specifically, the intention of feedback is to increase student understanding and performance in an identified domain of functioning (Shute, 2008). In the current educational context, the information contained within formative feedback can easily be delivered to learners via computers in online and/or blended classrooms with learning analytic capabilities (Long & Siemens, 2014). It is important to note however, that undirected computer administered feedback will not by itself produce effective teaching and learning. Thus, it is important to review past practice on how feedback is best personalized, as well as identify under what circumstances such feedback is successful.

One of the key findings within the feedback literature is the differential effect of personalized feedback on varying levels of student ability. In a study by Chen, Chang, and Wang (2008), Taiwanese students in a computer science course were treated with a ubiquitous electronic learning environment. The system would adapt its instruction and feedback practices based on the learner's previous performance in the course. While the entire treatment group performed on average greater than the control group, the key differences were in specific student subgroups. It was found that the higher performing students performed similarly between the control and treatment groups, while the lower performing students found benefit.

Similar results were found in Chen's (2008) study. Students in Chen's study were administered a pretest the at start of the course. This pretest estimates the students' ability level in specific content domains and creates a learning plan that adjusts their coursework accordingly. Students in the treatment group received a personalized learning path and feedback informed by their estimated subject ability. By contrast, the control group was able to freely pick and choose their own progression through the course content. The researchers found that the students receiving personalized feedback and instruction performed significantly higher on average, even when participants were matched on pretest scores (Chen, 2008).

Another theme in the literature surrounding personalized feedback is that of student satisfaction. Work by Gallien and Oomen-Early (2008) compared the effects of personalized feedback against collective feedback. Students in a health education course were provided with either of the two forms of feedback. The personalized group received feedback forms that had been filled out by the instructor personally. Such feedback included some corrective commentary, additional information, and some Socratic feedback that encouraged the student to reflect on their performance in the course. Students receiving the collective feedback received a form that summed up the entire class's performance on various assignments and projects, with no individuation. Results of the study indicated that the personalized group not only performed better on average in the class, but they reported higher levels of satisfaction with the course and their feedback (Gallien & Oomen-Early, 2008).

Building on the existing literature and introducing the idea of data informed feedback, Arnold and Pistilli (2012) examined the effectiveness of the Course Signals learning analytics dashboard on student academic performance. The dashboard used by Course Signals utilizes many student attributes and reports back to the student a set of lights, analogous to a traffic light. A green light indicates that Course Signals predicts the student will perform well in the course, while a yellow or red light imply that some intervention or greater effort on the part of the student is required to perform well or even pass. Though highly simplified, the dashboard is an effective and intuitive method of delivering formative feedback. Students who encounter yellow or red lights are encouraged to adapt their study skills in whatever way necessary to improve their situation. One of the most pertinent findings, and in congruence with Chen, Chang, and Wang (2008), is that the effectiveness of Course Signals was not uniform across the spectrum of student ability. Arnold and Pistilli (2012) found that high performing students did not benefit from the implementation of Course Signals. The rationale was that they already possessed the academic skills needed to perform well in a postsecondary setting. Further, it was found that the lowest performing group of students also did not benefit from the dashboard. It is suggested that these students lacked the motivation or skills required to turn their academic situation around. They found that the lower middle portion of the spectrum of ability was where the greatest benefit was observed. Course Signals made these students aware of their short-comings, and they used that feedback to improve their academic situation (Arnold & Pistilli, 2012)

The use of regular quizzing has also been used to inform and personalize feedback. A study by Pennebaker, Gosling, and Ferrell (2013) used an online quizzing system to deliver short tests to an introductory psychology class. Upon completion of the quizzes, students were given personalized feedback based upon their course performance, as the quizzes informed them of where they needed further study. Compared with previous years not using the system, it was found that the students performed significantly better. There was also a significant narrowing of the achievement gap between students of high and low socio-economic status. The authors theorized that the quizzing and feedback helped to promote student self regulation, as their grades in other courses being taken simultaneously also increased (Pennebaker, Gosling & Ferrell, 2013).

Work by Wright, Mckay, Hershock, Miller, and Tritz (2014) examined the effectiveness of E^{2}Coach, a learning analytic platform that delivers personalized advice feedback to students at the University of Michigan. Unlike the cases mentioned earlier, their study looked at whether students in the personalized group received grades higher than would otherwise be predicted by the system in a first-year physics course. They found that students in the personalized group achieved significantly higher grades than was predicted by the platform compared to those in the control group. Additionally, they found that there was a differential effect of the system with respect to gender, as men performed higher than women on average.

Kim, Jo, and Park (2016) conducted a study on the relevance of feedback satisfaction with learning analytics. The authors looked at the effect of a learning analytic dashboard on levels of satisfaction and course performance in a management statistics class. They found that on average the group allowed access to the feedback dashboard performed greater than the control group. By contrast however, the study found that higher academic achievers showed lower levels of satisfaction with the dashboard. Further, there was an overall negative relationship between satisfaction with the dashboard and usage (Kim, Jo & Park, 2016).

Building on the previous studies, the goal of this study is to assess the following hypotheses:

(1) Students who receive personalized feedback achieve significantly higher final grades than the control group.

(2) Students who receive personalized feedback show the greatest difference between final grades compared to the control group in the low middle portion of the distribution of student ability.

(3) Students receiving personalized feedback receive significantly higher final grades than predicted grades compared to the control group.

(4) Students who receive personalized feedback show the greatest difference in achieving higher than predicted grades in the low middle portion of the distribution of student ability.

(5) Students in the personalized feedback group display higher levels of feedback satisfaction compared to the control group.

Methods

Participants/Data

Study participants were undergraduate students from the University of Saskatchewan enrolled in two sections of a 100-level biology course. Between the two sections, total enrollment was approximately 800 students. After data cleaning procedures, the final sample consisted of 476 individuals (346 women, 130 men, average age = 18.8 years, age range = 17-50 years). Participants were asked at the beginning of the course if they would participate in the study. Students had the ability to decline participation in the study.

Demographic data was attained from the university's data warehouse. Students' electronic activity data was monitored and collected with the university's learning analytic platform Student Advice Recommender Agent known as SARA (Greer et al., 2015). An online survey was also administered to students at the end of semester to assess their satisfaction with the course, SARA, and many other aspects of the class. At the outset of the course, students were randomly assigned to one of two groups:

(1) Personalized Feedback Group – participants received personalized feedback through an email generated by SARA

(2) Control Group – participants received generic feedback through an email generated by SARA (this message was the same for all members of the control group)

SARA differentiated feedback based on 40 distinct student characteristics (e.g. student anticipated grade, student past performance in the course, postsecondary school experience, urban or rural background). Students received a new course feedback email every week; each new message utilized up to date information to adapt its feedback. For example, following the midterm exam, a student would receive a feedback message based on their midterm performance; should that student have not performed well, they may also be given advice on how to improve their mark (attending structured study sessions, increasing their study time and/or improving the quality of their studying). The control group also received a weekly message providing course feedback, though this message was common to all students within the control group.

Analytical Framework

In the present study, final grade refers to the grade that each student received in the course upon its completion and is measured out of a possible one hundred. The difference in predicted grade and final grade is determined by taking the student's final grade and subtracting their predicted grade (predicted grade based on SARA's algorithm) from it. Positive scores imply that the student achieved beyond what SARA predicted for them. Satisfaction was determined by score on a summated rating scale, with a maximum score of five. Students answered five questions pertaining to their sense of appreciation, belonging, encouragement, and personalization, and whether they thought SARA served as a good reminder of course responsibilities.

The statistical software program R (R Core Team, 2017) was used for data analysis. Student marks and corresponding survey data were analyzed using t-tests for independent samples to compare means for final grades, difference in predicted grades, and satisfaction scores. To examine the potential effect of gender, 2x2 factorial ANOVAs were also conducted on final grades, difference in predicted grades, and satisfaction scores.

Categories for student ability were determined by using students' first term GPA average. Four subsets of student groups were created according to their first term GPA. Ranges for the grade divisions were done in accordance with how SARA differentiates feedback (<51.75, 52.50<66.75, 67<81.40, and >81.6). To test hypotheses (2) and (4), the lower middle portion of the distribution was defined as those with a first term GPA ranging between 52.50 and 66.75. This approach is well founded in the literature (e.g., Arnold & Pistilli, 2012; Chen, Change, & Wang, 2008).

Results

Hypothesis 1

To assess the mean difference in final grade between the personalized feedback group and the control an independent samples t-test was used. With regards to final grade, Levene's test showed that variances were similar for both the personalized and control group F(1, 474) = .09, p > .05. It was found that on average the personalized group (M = 71.83, SD = 12.21) performed slightly better than the control group (M = 71.29, SD = 12.02) in terms of final grade. This difference was found not to be significant t(474) = -0.48, p > .05. The effect of gender was also assessed with reference to the above hypothesis with an ANOVA. No main effects were observed for either gender F(1, 472) = .61, p >.05, type of feedback F(1, 472) = .28, p > .05, or the interaction F(1, 472) = 1.54, p > .05. Hence, no differences or differential effects were observed by the inclusion of gender with respect to final grade.

Hypothesis 2

The mean differences in final grade between the personalized and control groups in the low middle portion were assessed with a Welch corrected two sample t-test as Levene's test revealed that variances for the personalized and control groups were not similar F(1, 113) = 4.68, p < .05. It was found that on average the personalized group (M = 60.30, SD = 8.00) performed better than the control group (M = 57.88, SD = 5.97) in terms of final grade. The difference was found to be not significant t(103.54) = -1.83, p > .05. An ANOVA taking into account gender was also conducted; results showed no main effects of gender F(1, 111) = .05, p > .05, type of feedback F(1, 111) = 3.25 p > .05, or their interaction F(1, 111) = .31, p > .05. Again, no differences were found when including gender as a factor.

Hypothesis 3

The difference between predicted grade at start of term and final grade across the personalized and control groups were assessed with an independent samples t-test. Levene's test showed that variances were similar for both personalized and control groups with respect to the difference in predicted grade and final grade F(1, 474) = .18, p > .05. It was found that on average the personalized group (M = 1.54, SD = 9.18) demonstrated higher than predicted final grades than the control group (M = .79, SD = 8.76). This difference was found not to be statistically significant t(474) = -.92, p > .05. An ANOVA was conducted to assess the potential influence of gender with grade difference and type of feedback. No main effect was observed for type of feedback F(1, 472) = .51, p > .05, or the interaction between type of feedback and gender

$F(1, 472) = 2.29$, $p > .05$. However, a main effect of gender was observed $F(1, 472) = 11.44$, $p < .001$, with a small effect size of $\omega^2 = .02$.

Hypothesis 4

With respect to the low middle portion of the distribution, the mean grade differences for both the personalized and controls groups were assessed with an independent samples t-test. Levene's test showed that variances were similar for both personalized and control groups $F(1, 113) = .88$, $p > .05$. Results showed that the personalized group ($M = -1.63$, $SD = 9.73$) performed better than the control group ($M = -2.86$, $SD = 10.27$) but the difference was found not to be significant $t(113) = -.66$, $p > .05$. An ANOVA was conducted to observe the effect of gender on the above variables. No main effect of type of feedback was observed $F(1, 111) = .24$, $p > .05$, nor was the interaction between type of feedback and gender found to be significant $F(1, 111) = 2.48$, $p > .05$. As with the previous analysis for the entire course, a main effect of gender was also observed $F(1, 111) = 5.12$, $p < .05$, with a small effect size of $\omega^2 = .01$.

Hypothesis 5

Satisfaction between the personalized and control groups was assessed with an independent samples t-test. Levene's test showed that variances were similar for both personalized and control groups with respect to satisfaction $F(1, 285) = 1.00$, $p > .05$. Results showed that the personalized group ($M = 2.74$, $SD = 1.54$) had higher levels of satisfaction with feedback compared to the control group ($M = 2.37$, $SD = 1.62$). This difference was found to be significant $t(285) = -2.00$, $p < .05$, with a small effect of $r = .12$. An ANOVA was conducted to examine the effect of gender on the satisfaction of feedback. No main effect was observed for gender $F(1, 283) = .89$, $p > .05$, but a main effect was observed for type of feedback $F(1, 283) =$

4.20, $p < .05$. More importantly the interaction between gender and type of feedback was also found to be significant $F(1, 283) = 6.18$, $p < .05$, with a small effect size of $\omega^2 = .02$. This result suggests that females and males showed different patterns of feedback satisfaction depending on which feedback group they belonged to. Simple effects analysis showed that the mean levels of satisfaction for females in the personalized group ($M = 2.96$, $SD = 1.40$) were significantly higher than females in the control group ($M = 2.27$, $SD = 1.64$). Cohen's d was computed as an effect size for the difference between these two female groups and it showed an approximately medium effect size ($d = .45$). No significant differences were found for males across groups.

Discussion and Conclusion

With new advances in data analytic technology, the ability to personalize feedback for individual learners within large online or blended classrooms is improving (Long & Siemens, 2014). It is worth noting however that the seamless delivery of computer adaptive feedback cannot by itself be considered good pedagogy. Therefore, more work is required to ensure that such systems provide feedback that is demonstrably beneficial in improving student achievement and learning. The present study aimed to provide evidence-based guidance on the parameters by which that feedback is administered. Statistical data, both descriptive and inferential, on the successes of a personalized feedback system with respect to final course grades, differences in predicted grades, and feedback satisfaction were presented. It was found that no significant differences existed between the personalized and control group with respect to both final grade and difference between final grade and predicted grade. Further, no significant differences were found between the personalized and control group with respect to final grade and difference between final grade and predicted grade for the lower middle portion of the distribution. A significant gender difference was observed with respect to the difference between final and predicted grade for both the entire sample and within the lower middle portion of the distribution. With regards to feedback satisfaction, it was found that the personalized group showed significantly higher levels of feedback satisfaction compared to the control group. It was also determined that females within the personalized group reported significantly higher levels of feedback than those in the control group, while males showed no significant difference.

The non-significant findings with respect to the first and third hypotheses are not necessarily unsurprising, as feedback research has been shown to contain a large body of inconsistent or sometimes contradictory findings (Shute, 2008). It also suspected that issues relating to sample size may have contributed to the non-significant findings. The differences between feedback groups in terms of final grade and predicted grade differences were not expected to be large and detecting such small differences would be better accomplished with much larger sample sizes. This is especially the case when considering the division of the larger sample into the subset grade categories. The sample of the lower middle portion of our distribution was approximately a third of the original sample, and such a reduction likely limited the statistical power of the analyses.

The finding of the gender difference with respect to predicted compared to final grades was interesting and unusual. It is possible that SARA systematically under predicts male performance, or that our sample contained a group of higher than usual achieving males. It is much more likely that additional information is needed to fully understand this finding. Wright et al.'s (2014) study observed that females used the E^2Coach learning analytics system significantly more often than their male counterparts. They also found that higher E^2Coach user females outperformed lower use females. Despite this, an achievement gap existed with males outperforming females on average. In the present study, net use of or total activity on SARA was not examined. Subsequent work will be sure to consider these variables and their relationships to our findings on gender.

Another reason for having so few significant differences may be the voluntary nature of the survey. Greater participation in the survey would have provided a more accurate picture of the students' attitudes and responses. Other limitations with the present study may have to do with the division of students into grade categories. Several other variables or methods were available for use in dividing the students into separate ability categories. The method chosen for the study was not done arbitrarily; however different group formation protocols may have resulted in subsets with subtly different statistical qualities and may have yielded different results. Limitations may also lie in the satisfaction scale used to determine a participant's overall satisfaction with SARA. The satisfaction scale was a post hoc creation, made by aggregating the scores across five pre-existing items that pertained to our conception of what satisfaction with SARA might mean. As such, it has not been formally assessed for evidence of psychometric reliability or validity. In direct relation to this point, the quality of the survey items may be another limiting factor for the study. The survey was not prepared with the present study in mind. It would have been preferable to have had reliable and valid survey items prepared in advance that aligned more closely with the aims of the study.

This paper supports the need for continued research in the area of personalizing feedback in the context of electronic learning environments. Results of this study may provide further direction in bettering the effectiveness of personalizing electronically administered feedback in similar university courses. It is our hope that this work will have contributed meaningfully to the goal of understanding what constitutes quality online education and feedback.

References

Arnold, K., & Pistilli, M. (2012). Course signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, (pp. 267-270).

Chen, C. M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers & Education, 51*(2), 787-814.

Chen, G. D., Chang, C. K., & Wang, C. Y. (2008). Ubiquitous learning website: Scaffold learners by mobile devices with information-aware techniques. *Computers & Education, 50*(1), 77-90.

Gallien, T., & Oomen-Early, J. (2008). Personalized versus collective instructor feedback in the online courseroom: Does type of feedback affect student satisfaction, academic performance and perceived connectedness with the instructor? *International Journal on ELearning, 7*(3), 463-476.

Greer, J., Frost, S., Banow, R., Thompson, C., Kuleza, S., Wilson, K., & Koehn, G. (2015). The student advice recommender agent: SARA. Saskatoon, Saskatchewan, Canada.

Kim, J., Jo, I.-H., & Park, Y. (2016). Effects of learning analytics dashboard: Analyzing the relations among dashboard utilization, satisfaction, and learning achievement. *Asia Pacific Education Review, 17*(1), 13-24.

Long, P. D., & Siemens, G. (2014). Penetrating the fog: Analytics in learning and education. *TD Tecnologie Didattiche, 22*(3), 132-137. doi:10.17471/2499-4324/195

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS One, 8*(11), 1-6. doi:http://dx.doi.org.cyber.usask.ca/10.1371/journal.pone.0079774

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. doi:10.3102/0034654307313795

Wright, M. C., McKay, T., Hershock, C., Miller, K., & Tritz, J. (2014). Better than expected: Using learning analytics to promote student success in gateway science. *Change: The Magazine of Higher Learning, 46*(1), 28-34.