Investigating the effect of automated feedback on learning behavior in MOOCs for programming

Hagit Gabbay
School of Education
Tel Aviv University
hg.astd@gmail.com

Anat Cohen
School of Education
Tel Aviv University
anatco@tauex.tau.ac.il

ABSTRACT

The challenge of learning programming in a MOOC is twofold: acquiring programming skills and learning online, independently. Automated testing and feedback systems, often offered in programming courses, may scaffold MOOC learners by providing immediate feedback and unlimited re-submissions of code assignments. However, research still lacks empirical evidence of their effect on learning behavior of MOOC learners, with diverse backgrounds and goals. Addressing this gap, we investigated the connections between the use of automated feedback system and learning behavior measures, relevant for MOOCs: engagement, persistence and performance. Further, two subjective measures of success are examined: sense of learning and intention fulfilment. In an experimental design, we analyzed data of active learners in a Python programming MOOC (N=4652), comparing an experimental group provided with automated feedback with a control group that did not. In examining the effect of automated feedback, prior knowledge of programming and Python was considered. Empirical evidence was found for the relation between automated feedback usage and a higher engagement and better performance, as well as higher attendance in "active watchers" and "high-performed completers" clusters, obtained by cluster analysis. Learners reports on their experience with the automated feedback system supported these findings. Regarding the subjective measures of success, however, no difference was found between groups. Our study and the offered future research may contribute to the considerations regarding the integration of automated feedback in MOOCs for programming.

Keywords

automate feedback, MOOC for programming, learning behavior, prior knowledge, educational data mining, cluster analysis

1. INTRODUCTION

Massive Open Online Courses (MOOCs) for programming have the potential to teach programming to a broad and diverse audience [28]. The high demand for computer professionals and labor market needs have led to an abundance of courses, with large numbers of enrolees [25]. However, many learners struggle in these courses. Learning programming is challenging, but MOOC learners face additional difficulties, as they have to self-regulate their learning (SRL) and to cope with course content almost without the assistance of instructors [11]. The provision of feedback may assist learners with these challenges. Feedback is considered essential in online learning, both as formative assessment, promoting learning and increasing learner engagement and as an acceleration of the SRL process [16].

In programming courses, the practice of writing and running code is the basis for acquiring software developing skills [5, 48]. An immediate, detailed and accurate feedback makes practice more effective and may significantly improve learning [38, 40]. But in large scale courses, and especially in MOOCs, instructors cannot provide feedback on each submission [15]. Automated testing and feedback (ATF) systems address the need, allowing an unlimited number of learners and submissions [22]. Upon uploading a solution to the system, the learner receives immediate feedback and is given the option to resubmit a revised code, to complete the learning process. Feedback may address syntax errors, the correctness of results, the efficiency of the code and whether it fulfils instructions accurately [13, 48]. It can consist of basic feedback, including only correct/incorrect information or presenting the correct answer, or a detailed feedback, suggesting possible causes of error and hints for the solution [22, 44].

ATF systems were developed decades ago, and today there is a variety of tools and systems employing diverse technologies and methods for testing and generating feedback [9, 22, 37]. Previous research suggested that incorporating an ATF system into a programming course improves affective measures, such as satisfaction and sense of learning [4, 6]. The automated feedback is perceived by learners as enhancing learning as well as contributing to motivation and engagement [2, 32, 37]. Results regarding the system's impact on performance, however, were not conclusive (e.g. [7, 13, 14]).

Yet, most studies in the field of ATF have only focused on frontal courses, or online courses offered as part of an academic curriculum. Students in these courses have extensive interaction with the faculty, which enhances their learning [36, 42] and might "overshadow" the impact of automated feedback on learning outcome [15]. In MOOCs, ATF system may have a different effect. Furthermore, the diverse goals and intentions of MOOC learners may affect their learning behavior and performance [24, 29], and consequently the impact of automated feedback [34].

To gain the full potential of ATF in MOOCs, empirical research needs to be conducted in order to better understand the impact of automated feedback on learning behavior and outcomes. Yet, empirical research in this field is still lacking [1]. With the aim of addressing the gap, this study set out to investigate the connections between automated feedback usage and learners' learning behavior and outcomes, in MOOCs for programming.

It is now well established from a variety of studies that the relevant measures for learning behavior in MOOCs are engagement (measured through learning activity), persistence and performance [18,

H. Gabbay and A. Cohen. Investigating the effect of automated feedback on learning behavior in MOOCs for programming. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 376–383, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853125

20, 24, 51]. Additional success measures refer to learners' perception of achieved knowledge levels and intention fulfilment [17, 39]. Hence, the research questions we posed are as follows:

RQ1: What are the differences, if any, in engagement, persistence, and performance between ATF users and not-users?
RQ2: Is there any difference in learners' perceptions of achieved knowledge levels and intention fulfilment between ATF users and not-users?

Previous studies suggested that prior knowledge, related to the course content, may influence learning behavior and the impact of automated feedback [28, 33, 46]. Therefore, the investigated learning behavior measures were analysed with prior knowledge as a covariate.

Using a quantitative method and learning analytics approach, we compared data of two groups of learners in a Python programming MOOC incorporating ATF system: An experimental group who was given feedback on the code assignments by using the system, and a control group who did not. Collecting data from the course and system logs, cluster analysis was applied to identify and compare learning behavior patterns. Subjective success measures were compared based on learners' self-report data.

2. RELATED WORK

2.1 Measures of learning behavior in MOOCs

Previous research has established that given the heterogeneity of learners in MOOC, the appropriate measures of learning outcomes differ from those of formal education context [10, 24]. The most commonly used indicator to measure learning outcomes in MOOCs is learning engagement, measured by the number of lecture videos watched, the number of posts to forums, the number of quizzes taken, and the number of tasks completed [20, 52]. Another common measure is persistence, as measured by the learner's determination to complete tasks and the achieved degree of progress. The relevant measures are therefore the number of attempts to solve course exercises, the number of learning units that were studied (watching video or attempt to solve an exercise) as well as the most advanced unit that was studied [10, 20]. Learner's performance in the course is often measured by the scores on the exercises and assignments [18, 47].

Several studies applied cluster analysis to identify learning behavioral patterns and classified learner' groups by similar learning characteristics. A key study, establishing the main characteristics of MOOC learners, suggested four identified groups: completers (students who completed most assignments), auditing (students who did few-to-no assignments but engaged in watching videos), disengaging (students who did assignments early in the course, but later stopped participating), and sampling (students who watched videos only in the beginning of the course) [24]. A similar research described 3 groups of MOOC learners as: (1) active engaged (those who submit assignments and were actively involved in forums); (2) passive engaged (those who watch video or show passive involvement in forums); and (3) disengaged (learners whose activity decreased throughout the course) [41]. Recently, a study investigated learners' interactions with video, exercises and discussion forums, identifying seven patterns of learners' behavior: tasters, downloaders, disengagers, offline engagers, online engagers, and two patterns characterized by moderate and high social engagement [21]. The variables used to derive these patterns were the number of videos watched, in-video questions answered, exercises and assignments submitted, thread views and activity in the discussion forums.

Considering the suggested measures of learning behavior in MOOCs, in the current study we first applied cluster analysis to identified groups of learners who behaved similarly and then investigated the connections between behavior patterns and ATF usage.

2.2 Subjective measures of success in MOOCs

A variety of reasons motivate learners to enroll in MOOCs [29], with a variety of learning outcomes to expect [39]. Thus, it has been proposed that the success of learning in MOOCs should be evaluated through learner-centered measures. One suggested criterion is fulfillment of learner intentions [17]. In research conducting self-reported surveys, learners' intention fulfillment was found to be correlated with their engagement in course activities [39]. "Sense of learning" or "sense of achievement" , representing the perceived increase of knowledge, is another criterion for successive learning, being in use in previous studies to measure learning success and outcomes in MOOCs [30, 49, 51].

2.3 The effect of ATF in MOOCs on learning

In the context of MOOCs for programming, only a few studies have examined the impact of automated feedback, provided on code assignment, on learning behavior and outcomes. Regarding the aforementioned affective measures, several studies noted that automated feedback was perceived by learners as improving performance, increasing engagement and having a positive impact on learning strategies [8, 27]. An interesting study suggested that learners who "committed" by formal registration to use an ATF system in MOOC for programming were more engaged in solving code assignments, in compared with those who used it only partially, without registration [12]. Recent study came up with similar results [43]. No effect of using the system was found, however, regarding performance and course completion rates.

Most studies on automated feedback propose advanced algorithms and new approaches to improve error detection and feedback accuracy, but do not evaluate its effectiveness as a learning tool [31, 45, 50]. Several research reports on future intention to evaluate the impact of the examined ATF system on learning outcomes, yet to be done [5, 25]. Others suggest factors to consider while choosing or developing ATF systems for MOOCs, but no empirical results are provided [48].

Overall, there seems to be some evidence to indicate that automated feedback has the potential to support learners and enhance learning success in MOOCs for programming. The findings of the current study contribute to empirically based knowledge in this area.

3. METHOD

3.1 Course and ATF system

To answer research questions, we conducted an experiment in MOOC to learn the Python programming language, offered on Edx-based platform for MOOCs. The course consists of nine learning units, each of which includes content videos with comprehension questions, closed exercises (such as multiple-choice questions), and code-writing assignments. Answering the closed exercises is followed by feedback (correct/incorrect and an explanation).

The ATF system integrated in the course is INGInious - an open-source software, suitable for online programming courses, providing grades and textual feedback for code assignments (for more details on INGInious, see [19]). The system was incorporated into the course as an external tool, and registration was necessary for access. It was configured to allow unlimited submission of solutions. The textual messages provided as part of the feedback were adapted to the code assignments, containing different levels of

feedback according to error-type (e.g. correct / incorrect, expected correct answer or more elaborated feedback), as classified by [44].

3.2 Experimental design

Using the cohort-mechanism embedded in Edx platform, we randomly divided the learners enrolled for the course into control group (control-g) and experimental group (atf-g). Learners in the experimental group gained access to the ATF system. Those who chose to use it uploaded solutions for the code assignments, received feedback, and then were able to resubmit revised solutions. In the current experimental set, however, it was not possible to get information about how the learners in the control group solved the code assignments.

3.3 Data resources and definitions of measures

Research questions were answered by gathering data from different sources and harvesting measures to be compared. Definitions of research variables are provided in this section.

1) Demographics and prior knowledge (PK) in programming and Python obtained by pre-course questionnaire. To avoid subjective assessment, PK was defined in a Boolean manner (there is / there is no PK). From learners' responses we derived three PK categories: None, Programming (other language), Python [3].
2) Learning behavior data, consists of engagement, persistence and performance, obtained from course event logs (table 1). As we do not have information regarding the number of code assignments solved by learners who did not use the ATF system (control group and learners in atf-g who chose not to use the system), it was not defined as a measure of activity.
3) Log files of the ATF system, which contain information on submitted solutions, were analyzed to assess the use of automated feedback (relevant only to the experimental group).
4) Subjective measures were collected from the "learning experience" questionnaire, completed by learners at the end of learning. "Sense of learning" refers to the learner's evaluation of the level of knowledge achieved at the end of learning process. Learners were asked to choose one of four statements, representing four levels of knowledge. Intention fulfilment defined as one of three values (Yes / Partially / No).
5) Learners in the experimental group were asked to answer two additional questions regarding their perception of how the use of the system affected their engagement (Likert scale 1-5): "The system contributed to the motivation to complete more tasks in the course", "The option to correct my solution and resubmit prompted me to make an effort for a higher score". The obtained responses were used as supportive data to the results of the comparison between the two research groups.

The research was conducted under the rules of ethics, while protecting privacy and maintaining the security of information, and in accordance with the approval of the university ethics committee.

3.4 Research population

The study was conducted in the second half of 2021. Following the screening of non-active enrollees and learners who did not provide information about prior knowledge, N=4652 learners were included for our final study population, 15.6% (724) of which in the experimental group (hereinafter: atf-g) and 84.4% (3928) in the control group (hereinafter: control-g). The imbalance between the research groups was caused by technical reasons, as the system was integrated for the first time during the course cycle in which the study was conducted. To ensure that none of this affected the results, all the analyses were repeated several times, comparing the experimental group (atf-g) with random groups drawn from the control group, of the same size as atf-g. Upon completion of this study, which was also a technical pilot, all learners in the following course cycles were able to enjoy the benefits of using the ATF system.

Table1 : Learning behavior calculated measures

	variable	Description
Engagement (activity)	watched video	Number of watched videos (0-29)
	watched units	Number of units in which at least one video was watched (0-9)
	active-watched ratio	Ratio between the number of videos in which the learner solved a comprehension question and the total number of videos watched (0-1)
	solved exercises	Number of closed exercises (CE) learner attempted (0-39)
	solved units	Number of units in which at least one exercise was attempted (0-9)
	mean attempts	Mean attempts per exercise
persistence	units touched	Number of units in which the leaner watched a video or attempted an exercise (0-9)
	max unit touched	The most advanced unit in which the learner watched a video or attempted an exercise (1-9)
performance	grade	The mean score in closed exercises (0-10)

Participants were all Hebrew speakers, which is the language in which the course is taught, and were interested in learning Python. In terms of gender identification, learners' distribution was 73.5% male, 25.9% female, and 0.6% unidentified. The learners ranged in age from less than 11 to over 75, with the majority in the age at the range of 12-34 (79.3%).

According to self-reported prior knowledge, 30.4% of learners had prior programming skills but not Python (PK=Programming), 15.1% reported of Python (and programming) knowledge (PK=Python) and 54.5% had no prior knowledge relevant to course content (PK=None). Chi-square test indicated no significant difference between the experimental and control groups regarding the distribution of prior knowledge.

4. FINDINGS

This section presents the results aimed to answer research questions. To control the impact of PK on the effect of automated feedback as a covariate, the comparison of the two research groups in all tests was conducted separately at each level of PK. The findings are presented without regard to this covariate unless it is found to affect the results in a significant manner.

In regard to the imbalanced research groups, five repeats of the analyses with random subgroups from control-g, equal to the size of atf-g, yielded identical answers to research questions. Therefore,

we have chosen to present the comparison between the experimental group ($N_{atf\text{-}g}$ = 724) and the entire control group ($N_{control\_g}$ = 3928).

Table 2: Descriptive statistics of learning behavior measures (Natf-g = 724, Ncontrol-g = 3928 M=mean, MD=median, SD=standard deviation)

variable	atf-g			Control-g		
	M	MD	SD	M	MD	SD
watched video	12.7	11	9.0	8.8	5	8.6
watched units	4.3	4	2.9	3.1	2	2.7
active-watched ratio	0.4	0.4	0.2	0.3	0.3	0.3
solved exercises	15.2	12	12.9	8.3	1	11.8
solved units	4.3	4	3.1	2.4	1	2.9
mean attempts	2.4	2.1	2.0	1.7	1.6	2.0
units touched	4.7	4	3.0	3.4	2	2.8
max unit touched	4.8	4	3.0	3.7	2	3.0
grade	0.8	1	0.3	0.5	0.8	0.5

4.1 RQ1: The connection between ATF usage, learning behavior and performance

4.1.1 Comparing learning behavior variables

The operational variables of learning behavior and performance (table 1) were calculated from data and compared between research groups. Mann-Whitney test was applied, as the homogeneity assumption required for the t-test was not met. In the experimental group, the mean and median values of all behavioral variables were higher, as illustrated in table 2 and figure 1. This difference was found to be significant at the p<.001 level (Mann-Whitney U ranged between 1050975 – 954295.5), with small-medium effect size, given by the rank biserial correlation (0.228 - 0.402).

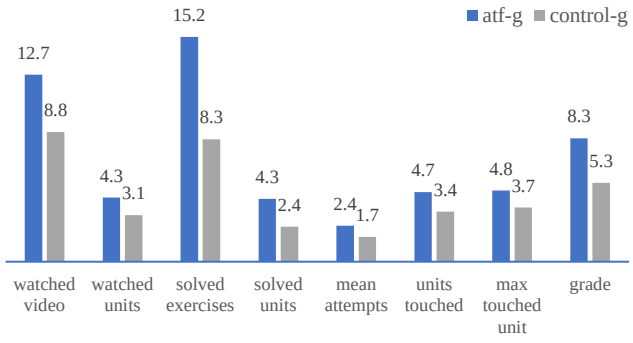

Figure 1: Mean values of learning behavior variables

Another way to measure persistence is to compare, per number of "touched" units, the percentage of learners from each group who reached this number. Chi-square test revealed significant difference between learners' percentage of each group (X^2(8) = 165.34, N = 4652, p<.001). From three units above, the number of units "touched" by the experimental group is higher (see figure 2).

More specifically, it can be claimed that a higher percentage of learners from the experimental group completed the course, i.e. learned all the lessons. Further evidence of this fact comes from a finding that a higher percentage of learners from the experimental group studied units 7-9, the advanced units of the course. A significant difference was indicated between learners' percentage from

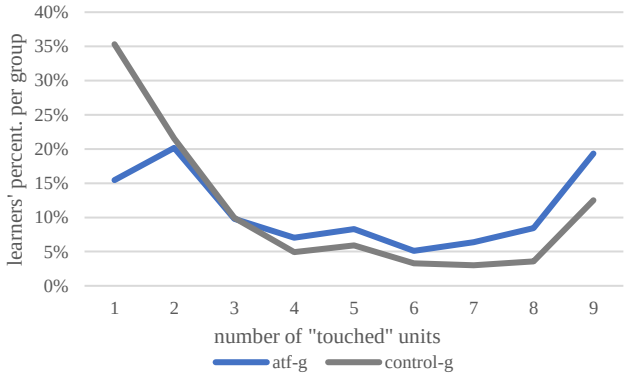

Figure 2: Comparison of learners' percentage per number of units "touched" (Natf-g = 724, Ncontrol-g = 3928)

each group who solved exercises in these units ($X^2_{unit\ 7}$ (2) = 100.00, $X^2_{unit\ 8}$(2) = 67.20, $X^2_{unit\ 9}$(2) = 34.42, p<.001). Nevertheless, for learners with prior knowledge of Python, the difference between the groups regarding solving the exercises of the last unit (unit 9) was not statistically significant.

4.1.2 Cluster analysis of learning behavior variables

Among the variables representing learning behavior and performance, five "differentiating" variables were identified using PCA: active watched ratio, solved units, mean attempts, max unit touched and grade. k was assumed a priori to be between 4 and 6 so that the clusters would be distinct, but not too many. The elbow method plot and silhouette score were then used to identify the most fitting number of five clusters, explaining 72.7% of variance in data (R^2 = 0.727). Table 3 summarizes the clusters obtained and the mean values of the differentiating variables within each cluster. Following the behavior patterns characterizing each cluster, they were named as follows: (1) "Touched and left": those who log in but showed almost no engagement with course content. This pattern, which represents learners who actually dropped out shortly after they started, was the most frequent. (2) "Completers, high performers": learners with highest performance and completing rates, while only moderate consumption of content. This pattern was the second in number of learners (3) "Active-watchers": those who watched video actively, answering in-video comprehension questions. (4) "Good starter, mean progress": those who solved correctly few exercises, mainly of the first one or two units, but had no intention to complete the course. (5) "Trail-error solvers": those who try to solve few exercises, with many attempts, low success and no progress. This was the least desirable behavior pattern.

Examining the presence of learners from each group within each of these clusters, presented in figure 3, revealed relatively high percentage of atf-g learners in clusters 2,3 (42.40%, 24.17%, respectively) and higher percentage of the control-g learners in cluster 1 (41.42%). Chi-square test indicated a statistically significant difference between atf-g and control-g (X^2(4) = 277.208, N = 4652, p<.001) over all levels of PK. Yet, cluster 2 was found to be significantly different from the other four clusters in terms of PK, as determined by one-way ANOVA (F(4,4647) = 32.664, p<.001)

and Tukey's HSD test for multiple comparisons (p<.001 for all the comparisons of cluster 2 and other clusters).

Table 3: Cluster characteristics and mean value of learning behavior variables (In bold: characterizing attribute of each cluster)

Cluster	1	2	3	4	5
Size	1707	1161	897	756	131
percentage of N=4652	36.91%	25.10%	19.39%	16.35%	2.83%
Active-watched ratio	0.142	0.428	**0.604**	0.15	0.329
Solved units	**0.164**	**7.205**	2.557	**1.655**	1.473
Mean attempts	0.179	2.556	2.338	2.319	**9.152**
Max unit touched	2.17	**7.999**	3.198	2.298	2.183
Grade	< 1	**95.1**	87.4	**95.1**	67.4

4.1.3 Learners' perception of ATF effects

Two questions about the effect of using the system were answered by 126 learners. In their responses, learners indicated they believed that using the ATF system affected their engagement and performance. The majority of respondents agreed with the statements that the option to correct and resubmit prompted them to make an effort for a higher score (87%) and using the ATF system motivated them to be more engaged in solving course exercises (81%). PK level had no impact on learners' perceptions.

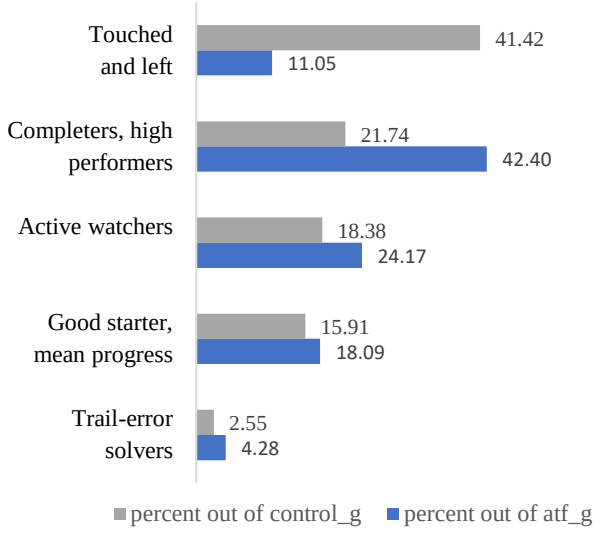

Figure 3: The percentage of learners from each group within each cluster

RQ2: Connections between automated feedback and success measures

Of the study population, 401 learners (9.87%) completed the "learning experience" questionnaire. Among the respondents, 32.27% (126) belong to the experimental group and 67.73% (275) to the control group. As can be seen in table 4, the majority of respondents described their achieved level of learning as "Can write basic code, needs more supported practice" (49.63%), followed by "Can write complex code and practice independently" (35.94%). Only 5.8% felt confident to start working in Python. To the question of intention fulfilments, however, the most frequent answer was "Yes" (81.91%).

Table 4: Subjective measures, learning experience questionnaire (N=401)

Sense of learning	
Understands the concepts, can't write code independently	8.56%
Can write basic code, needs more supported practice	49.63%
Can write complex code and practice independently	35.94%
Can work in Python	5.87%
Intention fulfilment	
No	0.98%
Partially	17.12%
Yes	81.91%

Applying the chi-square test to compare between research groups, no significant difference was found in all levels of PK. Notably, an expected significant dependency was found between PK and sense of learning ($X^2(6) = 67.73$, $N = 401$, p<.001), as prior knowledge led to higher assessment of the achieved level of knowledge.

5. DISCUSSION

With respect to the first research question, findings demonstrate connections between ATF system usage and learning behavior, as learners in the experimental group were more engaged with the course material and completed it at a higher rate. These results are in line with previous studies, suggesting that feedback (in general) enhances learners' engagement and persistence in online courses (e.g. [16, 20]) as well as in programming courses, not necessarily be offered online (e.g. [12]). Notably, in the current study all learners received feedback on the closed exercises, and the difference between groups was due to the additional feedback provided to the experimental group for code assignments. Therefore, it should be assumed that feedback on code assignments in MOOCs for programming is of utmost importance. Our findings, however, are contrary to those of [43] who did not observe a connection between feedback and learners' engagement. A possible explanation for this might be differences in feedback and course characteristics, which affect feedback effectiveness [33].

The current study indicates a connection between automated feedback and learners' performance and suggests positive trend of higher grades of the ATF users. Previous studies have not conclusively established this connection (e.g. [13, 14]). Nevertheless, due to our experimental design, we defined performance by closed exercises score and not by programming abilities (e.g. grades of code

assignments). As such, the evidence of higher performance of atfg in this case may indicate in fact a deeper understanding of programming principles [26].

Our clustering of learners based on learning behavior variables is similar to behavior patterns classified in previous studies [21, 24, 41], even though the number of clusters we selected is different. Learners of the experimental group were more "present" in the clusters described as "completers, high performers" and "Active-watchers" characterized by patterns identified as related to success in MOOCs [23]. The cluster analysis results support and expand the findings obtained for each variable as stand alone.

Interestingly, most of the analyses we conducted did not reveal any effect of prior knowledge on the connection between learning behavior and automated feedback. These findings do not support previous study, suggesting that prior knowledge affects the effectiveness of feedback in online learning environments [35]. Further research, with higher resolution determination of prior knowledge level, may shed light on this issue. One exceptional of our results is the case in Unit 9, where similar percentages of the Python experienced learners of two groups solved the closed exercises. It may be that the prior Python knowledge obscured the differences between groups, as experienced learners wanted to test their knowledge in this very specific unit, which is the most advanced one.

In support of the findings based on (objective) log files, the perceptions of the learners in the experimental group suggest that the automated feedback they provided heightened their motivation to be more engaged in solving exercises and encouraged them to score higher. This attitude is similar to learners' perception towards automated feedback reported in previous studies, in the context of in-class programming courses (e.g. [2, 32]). With regards to affective measures, therefore, automated feedback is perceived beneficial in both frontal and online learning environments.

With respect to the second research question, however, finding of the current study do not indicate a connection between the use of the ATF system and the subjective measures of success - intention fulfillment and sense of learning. These results differ from previous studies which have suggested that intention fulfillment is correlated with engagement in solving exercises in MOOC [39]. A two-way effect may have been created here: On one hand, the automated feedback may have led learners in the experimental group to more comprehensive learning [38] and on the other, they were more aware of errors and incorrect solutions, leading to a lower assessment of their abilities in Python.

6. CONCLUSIONS AND FUTURE WORK

Overall, the results of this study indicate a connection between the use of the ATF system, providing automated feedback, and learning behavior. Furthermore, the findings suggest that the automated feedback enhances the learners' engagement and persistence in the course as well as their performance. Nevertheless, we must be cautious in this context, and further research is needed to examine the effects of feedback on learning behavior (e.g. examining a "directional" connection). This is primarily due to the finding that the sense of learning and intention fulfilment were not affected by the use of ATF, suggesting the effect of feedback is likely to be complex and non-uniform within different facets of learning outcomes. However, the inability to obtain an objective assessment of the level of knowledge at the end of the learning is one of the current study limitations. The result of a final exam, for example, may reveal differences between research groups that are not apparent in self-reported evaluation. Nonetheless, there is no mandatory assessment in a MOOC due to its nature.

A further limitation is that the experimental design prevented a comparison of the research groups in regard to solving code assignments, which in fact is the subject of feedback. Future research be undertaken with a setup allowing the comparison of these data as well, might bring additional insight into the effect of automated feedback.

Lastly, the feedback provided by the ATF system was referred to in the current study as "black box". In light of the concept proposed by Narciss [33], which links the characteristics of feedback to its effectiveness, further experimental research is needed to analyze the effects of feedback characteristics (e.g. the structure of textual message) on learning in MOOCs for programming. Expanding empirical research knowledge regarding the impact of automated feedback on learning may contribute to the effective integration of ATF systems and thus promoting learners' acquisition of programming skills and achievement of learning goals.

7. ACKNOWLEDGMENTS

Our thanks to the Azrieli foundation for the award of a generous Azrieli Fellowship, which allowed this research. We are also indebted to our anonymous reviewers and the EDM program chairs for their valuable feedback.

8. REFERENCES

[1] Ahmed, U.Z., Srivastava, N., Sindhgatta, R. and Karkare, A. 2020. Characterizing the pedagogical benefits of adaptive feedback for compilation errors by novice programmers. *Proceedings - International Conference on Software Engineering* (2020), 139–150.

[2] Annamaa, A., Suviste, R. and Vene, V. 2017. Comparing different styles of automated feedback for programming exercises. *ACM International Conference Proceeding Series* (2017), 183–184.

[3] Bajwa, A., Bell, A., Hemberg, E. and O'Reilly, U.M. 2019. Student code trajectories in an introductory programming MOOC. *Proceedings of the 6th 2019 ACM Conference on Learning at Scale, L@S 2019* (New York, NY, USA, Jun. 2019), 1–4.

[4] Benotti, L., Aloi, F., Bulgarelli, F. and Gomez, M.J. 2018. The effect of a web-based coding tool with automatic feedback on students' performance and perceptions. *SIGCSE 2018 - Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (2018), 2–7.

[5] Bey, A., Jermann, P. and Dillenbourg, P. 2018. A comparison between two automatic assessment approaches for programming: An empirical study on MOOCs. *Educational Technology and Society*. 21, 2 (2018), 259–272. DOI:https://doi.org/10.2307/26388406.

[6] Cai, Y.Z. and Tsai, M.H. 2019. Improving Programming Education Quality with Automatic Grading System. *International Conference on Innovative Technologies and Learning* (Dec. 2019), 207–215.

[7] Cavalcanti, A.P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D. and Mello, R.F. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*. 2, (Jan. 2021), 100027. DOI:https://doi.org/10.1016/J.CAEAI.2021.100027.

[8] Chan, M.M., De La Roca, M., Alario-Hoyos, C., Plata, R.B.,

Medina, J.A. and Rizzardini, R.H. 2017. MOOCMaker-2017 Perceived Usefulness and Motivation Students Towards the use of a Cloud-based Tool to Support the Learning Process in a Java MOOC. *International Conference MOOC-MAKER* (2017), 73–82.

[9] Combéfis, S. 2022. Automated Code Assessment for Education : Review , Classification and Perspectives on Techniques and Tools. *Software 2022*. 1, (2022), 3–30. DOI:https://doi.org/https://doi.org/10.3390/software1010002.

[10] Evans, B.J., Baker, R.B. and Dee, T.S. 2016. Persistence Patterns in Massive Open Online Courses (MOOCs). *http://dx.doi.org/10.1080/00221546.2016.11777400*. 87, 2 (Mar. 2016), 206–242. DOI:https://doi.org/10.1080/00221546.2016.11777400.

[11] Feklistova, L., Luik, P. and Lepp, M. 2020. Clusters of programming exercises difficulties resolvers in a MOOC. *Proceedings of the European Conference on e-Learning, ECEL*. 2020-Octob, (2020), 563–569. DOI:https://doi.org/10.34190/EEL.20.125.

[12] Gallego-Romero, J.M., Alario-Hoyos, C., Estévez-Ayres, I. and Delgado Kloos, C. 2020. Analyzing learners' engagement and behavior in MOOCs on programming with the Codeboard IDE. *Educational Technology Research and Development*. 68, 5 (Oct. 2020), 2505–2528. DOI:https://doi.org/10.1007/s11423-020-09773-6.

[13] Gordillo, A. 2019. Effect of an Instructor-Centered Tool for Automatic Assessment of Programming Assignments on Students' Perceptions and Performance. *Sustainability*. 11, 20 (Oct. 2019), 5568. DOI:https://doi.org/10.3390/su11205568.

[14] Gusukuma, L., Bart, A.C., Kafura, D. and Ernst, J. 2018. Misconception-driven feedback: Results from an experimental study. *ICER 2018 - Proceedings of the 2018 ACM Conference on International Computing Education Research* (New York, New York, USA, Aug. 2018), 160–168.

[15] Hao, Q., Wilson, J.P., Ottaway, C., Iriumi, N., Arakawa, K. and Smith, D.H. 2019. Investigating the essential of meaningful automated formative feedback for programming assignments. *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC* (Oct. 2019), 151–155.

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

[17] Henderikx, M., Kreijns, K., Education, M.K.-D. and 2017, undefined 2017. Refining success and dropout in massive open online courses based on the intention–behavior gap. *Taylor & Francis*3 ,38 . (Sep. 2017), 353–368. DOI:https://doi.org/10.1080/01587919.2017.1369006.

[18] Hew, K.F. 2016. Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS. *British Journal of Educational Technology*. 47, 2 (Mar. 2016), 320–341. DOI:https://doi.org/10.1111/bjet.12235.

[19] INGInious [software] 2014. https://github.com/UCL-INGI/INGInious.

[20] Jung, Y. and Lee, J. 2018. Learning engagement and persistence in massive open online courses (MOOCS). *Computers and Education*. 122, (Jul. 2018), 9–22. DOI:https://doi.org/10.1016/j.compedu.2018.02.013.

[21] Kahan, T., Soffer, T. and Nachmias, R. 2017. Types of participant behavior in a massive open online course. *IRRODL*18–1 ,(2017) 6 ,18 .. DOI:https://doi.org/10.19173/irrodl.v18i6.3087.

[22] Keuning, H., Jeuring, J. and Heeren, B. 2018. A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education*. 19, 1 (2018), 1–43. DOI:https://doi.org/10.1145/3231711.

[23] Khalil, M. and Ebner, M. 2017. Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories. *Journal of Computing in Higher Education*. 29, 1 (Apr. 2017), 114–132. DOI:https://doi.org/10.1007/S12528-016-9126-9/FIGURES/7.

[24] Kizilcec, R.F., Piech, C. and Schneider, E. 2013. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *ACM International Conference Proceeding Series*. (2013), 170–179. DOI:https://doi.org/10.1145/2460296.2460330.

[25] Krugel, J., Hubwieser, P., Goedicke, M., Striewe, M., Talbot, M., Olbricht, C., Schypula, M. and Zettler, S. 2020. Automated Measurement of Competencies and Generation of Feedback in Object-Oriented Programming Courses. *2020 IEEE Global Engineering Education Conference (EDUCON)*.336–329 ,(2020) .

[26] Krusche, S. and Seitz, A. 2018. ArTEMiS - An Automatic Assessment Management System for Interactive Learning. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (New York, NY, USA, 2018), 284–289.

[27] Krusche, S. and Seitz, A. 2019. Increasing the Interactivity in Software Engineering MOOCs-A Case Study. *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019), 7592–7601.

[28] Luik, P., Feklistova, L., Lepp, M., Tõnisson, E., Suviste, R., Gaiduk, M., Säde, M. and Palts, T. 2019. Participants and completers in programming MOOCs. *Education and Information Technologies*. 24, 6 (Nov. 2019), 3689–3706. DOI:https://doi.org/10.1007/s10639-019-09954-8.

[29] Luik, P., Lepp, M., Feklistova, L., Säde, M., Rõõm, M., Palts, T., Suviste, R. and Tõnisson, E. 2020. Programming MOOCs–different learners and different motivation. *International Journal of Lifelong Education*. 39, 3 (May 2020), 305–318. DOI:https://doi.org/10.1080/02601370.2020.1780329.

[30] Magen-Nagar, N. and Cohen, L. 2017. Learning strategies as a mediator for motivation and a sense of achievement among students who study in MOOCs. *Education and Information Technologies*. 22, 3 (May 2017), 1271–1290. DOI:https://doi.org/10.1007/S10639-016-9492-Y/FIGURES/3.

[31] Marin, V.J., Pereira, T., Sridharan, S. and Rivero, C.R. 2017.

Automated personalized feedback in introductory Java programming MOOCs. *Proceedings - International Conference on Data Engineering*. (2017), 1259–1270. DOI:https://doi.org/10.1109/ICDE.2017.169.

[32] Marwan, S., Fisk, S., Price, T.W., Barnes, T. and Gao, G. 2020. Adaptive Immediate Feedback Can Improve Novice Programming Engagement and Intention to Persist in Computer Science. *The 2020 ACM Conference on International Computing Education Research*–194 ,(2020) 203.

[33] Narciss, S. 2013. Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review*. 23, 1 (2013), 7–26. DOI:https://doi.org/10.1344/der.2013.23.7-26.

[34] Narciss, S. 2008. Feedback strategies for interactive learning tasks. *Handbook of research on educational communications and technology*. J.M. Spector, M.D. Merrill, J. Van Merriënboer, and M.P. Driscoll, eds. Lawrence Erlbaum Associates, Mahaw, NJ. 125–144.

[35] Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G. and Melis, E. 2014. Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers and Education*. 71, (Feb. 2014), 56–76. DOI:https://doi.org/10.1016/j.compedu.2013.09.011.

[36] Nicol, D.J. and Macfarlane-Dick, D. 2006. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*. 31, 2 (Apr. 2006), 199–218. DOI:https://doi.org/10.1080/03075070600572090.

[37] Pettit, R. and Prather, J. 2017. Automated assessment tools: too many cooks, not enough collaboration. *Journal of Computing Sciences in Colleges*. 32, 4 (2017), 113–121.

[38] Pieterse, V. 2013. Automated Assessment of Programming Assignments. *3rd Computer Science Education Research Conference on Computer Science Education Research*. 3, April (2013), 45–56. DOI:https://doi.org/http://dx.doi.org/10.1145/1559755.1559763.

[39] Rabin, E., Kalman, Y.M. and Kalz, M. 2019. An empirical investigation of the antecedents of learner-centered outcome measures in MOOCs. *International Journal of Educational Technology in Higher Education*. 16, 1 (Dec. 2019), 1–20. DOI:https://doi.org/10.1186/S41239-019-0144-3/TABLES/4.

[40] Rafique, W., Dou, W., Hussain, K. and Ahmed, K. 2020. Factors influencing programming expertise in a web-based e-learning paradigm. *Online Learning Journal*. 24, 1 (2020), 162–181. DOI:https://doi.org/10.24059/olj.v24i1.1956.

[41] Ramesh, A., Goldwasser, D., Huang, B., Iii, H.D. and Getoor, L. 2013. Modeling learner engagement in MOOCs using probabilistic soft logic. *NIPS workshop on data driven education* (2013), Vol. 21, 62.

[42] Restrepo-Calle, F., Ramírez Echeverry, J.J. and González, F.A. 2019. Continuous assessment in a computer programming course supported by a software tool. *Computer Applications in Engineering Education*. 27, 1 (Jan. 2019), 80–89. DOI:https://doi.org/10.1002/cae.22058.

[43] Serth, S., Teusner, R. and Meinel, C. 2021. Impact of Contextual Tips for Auto-Gradable Programming Exercises in MOOCs. *Proceedings of the Eighth ACM Conference on Learning @ Scale* (New York, NY, USA, 2021), 307–310.

[44] Shute, V.J. 2008. Focus on Formative Feedback. *Review of Educational Research*. 78, 1 (Mar. 2008), 153–189. DOI:https://doi.org/10.3102/0034654307313795.

[45] Singh, R., Gulwani, S. and Solar-Lezama, A. 2013. Automated feedback generation for introductory programming assignments. *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*. (2013), 15–26. DOI:https://doi.org/10.1145/2462156.2462195.

[46] Smits, M.H.S.B., Boon, J., Sluijsmans, D.M.A. and van Gog, T. 2008. Content and timing of feedback in a web-based learning environment: effects on learning as a function of prior knowledge. *Interactive Learning Environments*. 16, 2 (Aug. 2008), 183–193. DOI:https://doi.org/10.1080/10494820701365952.

[47] Soffer, T. and Cohen, A. 2019. Students' engagement characteristics predict success and completion of online courses. *Journal of Computer Assisted Learning*. 35, 3 (2019), 378–389. DOI:https://doi.org/10.1111/jcal.12340.

[48] Staubitz, T., Klement, H., Renz, J., Teusner, R. and Meinel, C. 2015. Towards practical programming exercises and automated assessment in Massive Open Online Courses. *Proceedings of 2015 IEEE International Conference on Teaching, Assessment and Learning for Engineering, TALE 2015* (2015), 23–30.

[49] Stich, A.E. and Reeves, T.D. 2017. Massive open online courses and underserved students in the United States. *The Internet and Higher Education*. 32, (Jan. 2017), 58–71. DOI:https://doi.org/10.1016/J.IHEDUC.2016.09.001.

[50] Wang, K., Lin, B., Rettig, B., Pardi, P. and Singh, R. 2017. Data-driven feedback generator for online programing courses. *The Fourth (2017) ACM Conference on Learning@ Scale*260–257 ,(2017) .

[51] Wei, X., Saab, N. and Admiraal, W. 2021. Assessment of cognitive, behavioral, and affective learning outcomes in massive open online courses: A systematic literature review. *Computers & Education*. 163, (Apr. 2021), 104097. DOI:https://doi.org/10.1016/J.COMPEDU.2020.104097.

[52] Xiong, Y., Li, H., Kornhaber, M.L., Suen, H.K., Pursel, B. and Goins, D.D. 2015. Examining the Relations among Student Motivation, Engagement, and Retention in a MOOC: A Structural Equation Modeling Approach. *Global Education Review*. 2, 3 (2015), 23–33.