ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(Special Issue); 320-331

Examining Students' Online Formative Test-Taking Behaviors Using Learning Analytics

Alper BAYAZIT* Denizer YILDIRIM** Gökhan AKÇAPINAR*** Hale ILGAZ****

Abstract

In online learning environments, assessment is an important dimension and also one of the most challenging parts of the online learning process. So, to provide effective learning, analyzing students' behaviors is important when designing online formative and summative assessment environments. In this study, students' profiles were analyzed within an online formative assessment environment and compared with a summative assessment environment based on attempt count, overall time spent, first-attempt score, and last-attempt score metrics. The within-subjects design, cluster analysis, and the Kruskal Wallis-H Test was carried out for analyzing behaviors. As a result, it was shown in the data that there were three main clusters. Cluster 1 showed a high number of interactions, and an increasing trend was observed in "grades" over "attempts". Additionally, Cluster 2 consisted of students who received the best grades in all other clusters, and finally, Cluster 3 consisted of students who interacted little and scored lower on formative assessments.

Keywords: test-taking behavior analysis, learning analytics, formative assessment, assessment analytics

Introduction

Assessment is one of the primary components of the online learning process. Two of the main approaches used in assessment design are formative and summative assessment. While both assessment approaches are focused on student development and progress, their approaches differ. Harlen and James (1997) defined formative assessment as an "assessment for learning" that focuses on student learning at the current stage and supports learning for the next step. While summative assessment, also defined as an "assessment of learning", is considered a more systematic and continuous recording of overall achievement. As a result, formative assessment focuses on the improvement of learning, whereas summative assessment focuses on providing information for accreditation and evaluation (Xiong et al., 2018). In an online course based on a formative assessment approach, students are aware of their strengths and weaknesses, can be more engaged and motivated, and monitor their progress (Crisp & Ward, 2008; Wolsey, 2008). Additionally, formative assessment can help decrease students' anxiety levels (Cassady & Gridley, 2005) and increase interaction between peers and instructors (Vonderwell et al., 2007). For this reason, the design of formative assessment is important in terms of providing information about how students perceive this process.

* Asst. Prof. Dr., Ankara University, Faculty of Medicine, Ankara-Turkiye, abayazit@ankara.edu.tr, ORCID ID: 0000-0003-4369-587X
** Dr., Ankara University, Faculty of Open and Distance Education, Ankara-Turkiye, dyildirim@ankara.edu.tr, ORCID ID: 0000-0002-4534-8153
*** Assoc. Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkiye, gokhana@hacettepe.edu.tr, ORCID ID: 0000-0002-0742-1612
**** Assoc. Prof. Dr., Ankara University, Faculty of Open and Distance Education, Ankara-Turkiye, hilgaz@ankara.edu.tr, ORCID ID: 0000-0001-7011-5354

To cite this article:
Bayazıt, A., Yıldırım, D., Akçapınar, G. & Ilgaz, H. (2023). Examining students' online formative test-taking behaviors using learning analytics, *14*(Special Issue), 320-331. https://doi.org/10.21031/epod.1275597

Received: 2.04.2023
Accepted: 3.09.2023

Technology use during the formative assessment process is especially seen as a pillar of tracking students' learning and performance within 21st-century learning environments (Shin et al., 2022). However, due to the nature of the formative assessment process, students sometimes tend to put less effort into these tests, which can ultimately cause poorer results compared to the summative assessment process (Wise, 2006; Yildirim-Erbasli & Bulut, 2022). The effort shown during test-taking is an important dimension of the formative assessment environment and allows the documentation of student engagement (Wise et al., 2013; Yildirim-Erbasli & Bulut, 2020). Test-taking frequency (Blondeel et al., 2023; Palmen et al., 2015; Shin et al., 2022) and response time and patterns (Man et al., 2018; Yildirim-Erbasli & Bulut, 2020; Wise & DeMars, 2006) are also indicators of students' engagement and motivation within the formative assessment process.

It is essential to investigate students' behaviors during the self-assessment process to find patterns that have a negative effect on learning (Yang et al., 2022). Also, revealing test-taking behavior allows for the identification of students' trial-and-error patterns and cheating activities (Man et al., 2018) and provides timely feedback for maintaining mastery of learning (Hui, 2023). It also enables researchers and/or administrators to gain greater insight into the processes and behaviors that lead to a specific test outcome (Stadler et al., 2020). Investigation of formative assessment behaviors also helps to make intervention possible and can significantly impact student interests and achievements (Rakoczy et al., 2019). Typically, these scales have been employed by researchers to examine exam-taking behaviors; however, self-report measures are weak against many forms of bias, and at times these scales can be quite limited, as they only provide fundamental information regarding a student's motivation towards test taking. Additionally, it can be hard to know how accurately test takers complete the scale (Wise & Gao, 2017). Thus, log data analysis can provide important information regarding students' behavior in formative online assessments for both professionals and educators (Guo, 2021). This data can also be useful for detecting differences in students' aberrant behavior in real time (Han & Kang, 2021) because it is a reliable predictor of the ability to be tested (Stadler et al., 2020). For example, students' interactions during the evaluation process are typically stored in log data. Some of these metrics include the number of student attempts, number of question views, number of hints viewed, number of submissions (Yang et al., 2022), response time data, total scores (Guo & Ercikan, 2021), test scores and submission times (Hui, 2023), individual mean item response times (Lee & Haberman, 2016), omitted items (Sarac & Loken, 2022), time-on-task and the number of interactions (Stadler et al., 2020), and flagged, reviewed, changed, or omitted items (Wise & Gao, 2017). These studies show that the trial scores, response time, and trial numbers are the most often used metrics. However, the pre-processing of data obtained from several assessment tools is seen as a limitation, and as a result, obtaining the necessary metrics from commonly used learning management systems can contribute to the literature and provide important insights into students' test-taking strategies and student profiles.

One test-taking performance study by Silm et al. (2013) focused on performance in low-stakes tests. According to their aim, they specifically searched for the number of items test-takers attempted to solve, the number of correct answers, overall time spent, and the speed of their accomplishments. The research was focused on 327 first-year students attending a higher education institution. It was found that when the difficulty levels of items are similar, the number of items solved and the mean time for each item can predict performance in low-stakes tests and short response times signal low-test scores. It was also discovered that the mean time for incorrect answers was shorter than the mean time for correct answers. Additionally, in another study using responses and response times for computer-based reading assessment, Yildirim-Erbasli and Bulut (2020) evaluated the effect of students' test-taking efforts on their reading growth. A quick screening tool was designed and applied to 7602 students over an academic year to monitor and assess their reading ability. They found that rapid guessing and slow responses can be helpful when calculating and interpreting students' growth estimates. In a large-scale test, Programme for International Student Assessment (PISA) 2012, Lundgren and Eklöf (2020) focused on test-takers' within-item behaviors from a self-reported and behavioral effort perspective. Essentially, they analyzed time on task, time to first action, number of actions, unique routes, repeated wrong routes, and actions per minute variables with self-reported data by using math test scores. Thus, they determined that low levels of effort before completing a task may not be diagnostic of test-taking motivation,

although low levels of effort before taking a test appear to be below the level of effort put in prior to giving up on a task. So, these variables/metrics provide important information when analyzing and interpreting the formative assessment results.

Clustering is employed for profiling purposes due to the clustering technique's interpretable insights regarding the relationship between test administration decisions and student performance profiles (Shin et al., 2022). Profiling test-takers or profiling students' test-taking behaviors are two of the primary approaches in clustering studies. For example, Yang et al. (2022) used cluster analysis to analyze how students interacted during formative assessments, which were given as a post-class self-evaluation. In the results, three distinct student profile clusters were determined. The students in Profile 1 were those who engaged the assessment system and hinted at it sparingly. While in Profile 2, students participated in exams, reviewed questions, and struggled to answer them. On the other hand, students in Profile 3 were those who successfully remembered information and used the exam system most. Therefore, according to these findings, students who completed online tests after class typically scored higher than those who did not, and those who engaged in non-standard behavior during the test did not increase their performance. Tempelaar et al. (2018) demonstrated that different at-risk groups might be identified by clustering several interaction data items including formative assessments. They highlighted that appropriate interventions are available by identifying these profiles. Additionally, Guo (2021) found four distinct student profiles in online exams taken at home or testing centers. Test-takers in only certain clusters tended to spend the majority of their time solving items, whereas, in other clusters, they were to have read the exam instructions for a significant portion of the time. In another study, Stenlund et al. (2018) clustered test-takers into groups and discovered three distinct student profiles: moderate, calm risk-taker, and test anxious risk-averse profile. Furthermore, in group difference studies, it was revealed that in terms of test performance, the calm risk-taker profile was the most successful, while the test anxious risk-averse profile was the least successful. Another approach in profiling studies is clustering individual test attempts rather than clustering an average over modules or students. Hui (2023) also stated that possible differences between scores and test types could be clustered and analyzed separately. Thus, computer-based techniques can be utilized to detect students who do not exhibit normal patterns by revealing behavior patterns through clustering. For example, Liao et al. (2021) claimed it is challenging to identify "item harvesters" who memorize or share test items. Thus, they offered a two-stage solution to identify this behavior, which, in the end, appeared to make tests less reliable. As a result, the initial phase should include cluster analysis to identify learners' test-taking behaviors, and then, abnormal behaviors must be marked for further investigation.

The clustering technique provides valuable information within a learning analytics framework to detect these general behaviors. It was shown in the presented literature that especially test-taking effort, engagement, and motivation play important roles in online formative assessments. When focusing on these dimensions, researchers worked in particular with the number of items that test-takers attempted to solve, the number of correct answers, overall time spent, and response time. Importantly, these user metrics had not been previously found to be analyzed by comparing summative and/or online formative test environments. Therefore, in the current study, students' profiles were analyzed within an online formative assessment environment and then compared with the summative assessment environment based on the metrics of attempt count, overall time spent, first attempt score, and the last attempt score.

Research Questions:

1. How many different groups are students divided into according to the metrics of taking the formative test?
2. Is there a significant difference between students in different groups in terms of their summative test scores?

Method

In the current study, a within-subjects design was used, and comparisons were made of the test-taking behaviors of participants in the weekly quizzes (formative assessment) and mid-term exams (summative

assessment). Features related to the test-taking behaviors of students were extracted from the database through data mining methods. Also, cluster analysis and the Kruskal Wallis-H Test were used to test whether there was a statistically significant difference in test-taking behaviors of students for the online formative and summative tests.

Participants

This study was conducted at a state university with 66 vocational school students enrolled in Web Programming II courses as part of the Computer Programming and Internet and Network Technologies Distance Education Programs. The students' ages ranged from 18-45 years old, and the group included 51 males and 15 females. The duration of the course lasted approximately 14 weeks, and the theoretical transfer and implementation of the course were conducted online. Finally, the mid-term and final exams in this study were administered in a face-to-face setting.

Procedure

The research process was carried out between the sixth and eighth week of the course. Students had access to the formative assessment test following the lecture in week six until week eight when the mid-term exam was held. This process involved the formative assessment design that included the processes of "finding and handling information" and "assessment" in the Learning Design Taxonomy (Toetenel & Rientes, 2016). The formative test was structured using a question bank containing questions that had previously been used for the mid-term exam in the previous semester. Importantly, the question bank included subjects relevant to the mid-term examination and was systematically categorized into distinct units. Notably, the distribution of questions across units within the question bank was found to be uneven. Nonetheless, during the examination process, each student uniformly encounters an equivalent number of questions from each unit. To illustrate, each student has posed four questions sourced from both Unit 1 and Unit 2. These questions might manifest either as a duplicate or exhibit some variation amongst themselves. In each attempt of the formative test, the students were randomly assigned 20 questions from the question bank. The students were then asked to answer the questions they had been provided within 30 minutes. At the end of each attempt, the question itself, the students' answers, and feedback regarding the correct or not correct answers were shown. Furthermore, the students were not shown the correct answers to any of the questions. When the students did answer a question incorrectly, they were expected to peruse the course resources themselves to determine the correct answer as well as review the instructor's lecture recordings and discuss what they had learned.

The summative assessment was comprised of a total of 20 multiple-choice questions, and all students engaged in the examination under uniform conditions. Additionally, the examination was administered according to a specific online framework. Also, students were diligently supervised by an instructor through an online live virtual classroom. Furthermore, the students' interactions within the web browser were subject to scrutiny via the implementation of a Proctoring Moodle Plugin. As a result, following the conclusion of the examination, analysis of the test items was carried out and indices related to the item difficulty spanned a spectrum from 0.23 to 0.78. Therefore, the reliability coefficient of the items, as assessed through Cronbach's alpha measure, was determined to be 0.695.

Data Analysis

In answering the first research question, students' formative exam-taking behaviors (digital traces during the exam) were analyzed using the four metrics mentioned in Table 1. Cluster analysis was used to group the students based on their exam-taking behaviors, and with this analysis, it was determined there were common exam-taking behaviors among the students. Due to there being no prior insight into the number of clusters, the number was determined automatically using the Silhouette metric. Thus, the k-Means algorithm in Orange data mining software was used for the cluster analysis. Since the distributions of the data were not in a standard range, normalization was applied in the data pre-processing. Additionally, the Silhouette scores of clusters from 2 to 9 were calculated to determine the ideal number of clusters, and the number of clusters with the highest Silhouette score was considered as the ideal number of clusters.

Table 1.
Metrics used in the clustering and their explanations

Metric	Description
Attempt count	Total number of attempts by the student in the formative assessment activity
Overall time spent	Total time spent by the student in formative assessment activities (min)
First attempt grade	Student's grade for the first attempt in the formative assessment activity
Last attempt grade	Student's grade for the last attempt in the formative assessment activity

Next, to answer the second research question, a Kruskal Wallis-H test was used to compare summative test scores based on clusters. During this process, a non-parametric statistical analysis was also used due to the normality assumption not being met.

Results

In this study, students' profiles were analyzed within a formative assessment environment and then compared with an established summative assessment environment by gathering several learning management system metrics. Thus, these metrics were used for clustering students and determining their test-taking behaviors. As a result, to analyze this situation in further depth, two research questions were studied.

RQ1. How many different groups are students divided into according to the metrics of taking the formative test?

When the Silhouette scores of clusters 2-9 were examined, it was recognized that the students were ideally divided into three clusters. When considering the cluster centroids provided in Figure 1, along with the descriptive statistics provided in Table 2, it was determined that the students in Cluster 3 (n = 7) made fewer attempts (Mdn = 2) and spent less time (Mdn = 10 min) in the formative assessment activity than students in the other cluster. Additionally, when the first and last attempt grades were examined, it was noteworthy that both were found to be low, and there was a negligible score increase within students' last trial compared to their first trial.

Next, students in Cluster 1 (n = 20) were those who made the most attempts (Mdn = 11) and also spent the most time in the formative assessment activity (Mdn = 194 min). As a result, according to the median trial numbers presented in Table 2, students within this group made four times more attempts than those in Cluster 2 and six times more than those in Cluster 3. Similarly, considering the median time spent, students in Cluster 1 spent four times more time than those in Cluster 2 and 20 times more than those in Cluster 3. However, when their scores from the exams were examined, it was recognized that their first-attempt grades were low, and their last-attempt grades were seen to be high. In other words, it was observed that there was a three-fold increase between their first and last attempt grades.

On the other hand, students in Cluster 2 (n = 38) were found to fall in between Cluster 1 and Cluster 3 in terms of the number of attempts (Mdn = 3) and time they spent (Mdn = 50 min) when considering their cluster centroids. However, despite fewer attempts, their first-attempt grades were determined to be higher than both clusters (Mdn = 15). This value was two times higher than the median first-attempt grades of Cluster 1 and approximately three times higher than that of Cluster 3. Therefore, it was determined there was an increase of four points between the median grades of their first and last attempt.

Figure 1.
Cluster Centroids

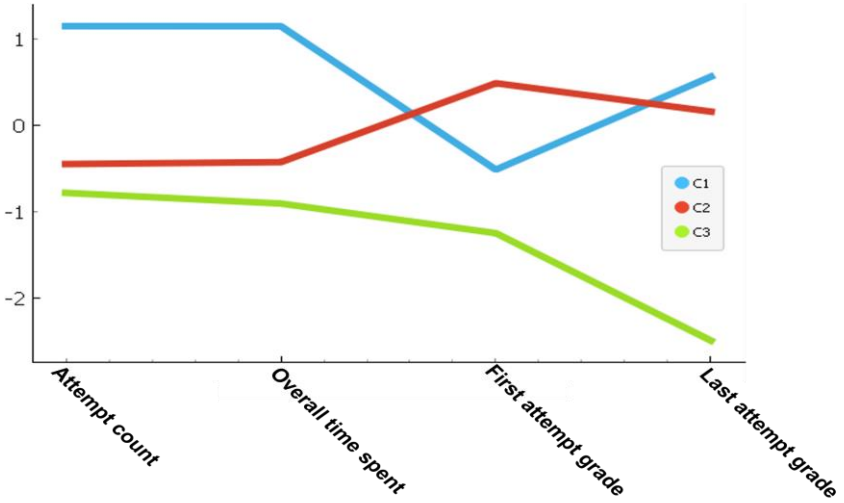

Table 2.
Descriptive statistics by clusters

Cluster	Statistics	Attempt count	Overall time spent	First attempt grade	Last attempt grade
Cluster 1	Mean	12.15	205.52	5.35	17.05
	N	20.00	20.00	20.00	20.00
	Std. Deviation	6.11	100.84	3.84	2.42
	Median	11.00	194.02	5.50	17.00
Cluster 2	Mean	3.28	55.18	10.69	15.00
	N	39.00	39.00	39.00	39.00
	Std. Deviation	1.88	36.19	4.68	2.47
	Median	3.00	50.30	11.00	15.00
Cluster 3	Mean	1.43	9.36	1.43	1.57
	N	7.00	7.00	7.00	7.00
	Std. Deviation	.53	9.40	2.51	2.70
	Median	2.00	9.60	4.00	5.00

RQ 2. Is there a significant difference between students in different groups in terms of summative test scores?

When comparing the summative test scores according to clusters, the non-parametric Bonferroni Correction with the Kruskal Wallis-H Test was applied since Levene's Test statistic was significant (L = 5.890, df1 = 2, df2 = 59, p <. 01).

Table 3.
Kruskal Wallis-H Test results

Cluster[a]	**Mean**	**N**	**SD**	**Median**	**Mean Rank**	**Chi-Square[b]**	**df**	**p**
Cluster 1	35.500	20	11.34	35.00	23.05			
Cluster 2	49.359	39	18.36	40.00	36.67	8.747	2	.013[c]
Cluster 3	33.333	3	5.77	30.00	20.67			
Total	44.113	62	17.28	40.00				

a. Kruskal Wallis-H Test
b. Grouping Variable: Cluster, Dependent Variable: Summative Grade
c. Significant: Bonferroni Correction $\frac{\rho}{n(n-1)/2}$ = *.05/3 = .01667* ⇒ .013 < *.01667*

According to the results from the Kruskal Wallis-H Test (Table 3), it was determined that there was a significant difference between clusters in terms of summative test scores (Chi-Square = 8.747; df = 2, < .0167). Thus, learners in the high-scoring group with fewer attempt counts in the formative test (Cluster 2, Mean: 49.359, Mean Rank: 36.67) achieved higher success within the summative test than those from the other two clusters. As a result, a significant difference was found to be present only between Cluster 1 and Cluster 2 in terms of summative assessment scores (U = 220, W = 430, Z = -2.734, *p* = .006, Adj. *p* = .016). Furthermore, no significant difference was found between Cluster 3 - Cluster 1 (U = 29, W = 35, Z = -0.93, *p* = .830, Adj. *p* = 1.00) and Cluster 3 - Cluster 2 (U = 27, W = 33, Z = -1.545, *p* = .137, Adj. *p* = .411).

Discussion

Analysis of learning progressions in exams, along with the dimensions of task design, trustworthiness, and fairness, is one way to take advantage of the potential of assessment analytics (Gašević et al., 2022). In the current study, the formative exam metrics collected from distance education students and students' test-taking behaviors in formative assessment were investigated through cluster analysis. As a result, it was revealed through the cluster analysis that three distinct student profiles were present via the assessment analytics.

Additionally, students in Cluster 1 showed a high number of interactions; while receiving low scores in their first attempts, they increased their scores in further attempts. When considering the first attempts, these students exhibited low achievement, but after making some effort, their scores improved. Furthermore, as a strategy, they may have reviewed the questions and then attempted to solve them by taking the exams again. Liao et al. (2021) discovered a similar behavior pattern in their studies and mentioned them as "item harvesters" who tended to remember, record, and then share items included in the test among their peers. Interestingly, these students caused security concerns within the high-stakes exams. On the other hand, the students could have discovered this pattern for themselves within the formative assessment process. Hui (2023) also determined similar patterns, stating some students exhibited developing patterns for discovering potential correct answers. In the current study, students in Cluster 1 made several attempts, received low scores on their first attempts, and then increased their scores in subsequent attempts. As a result, this may be evidence of a pattern related to formative assessment item harvesting via memorizing test items along with options for further attempts. Importantly, possible reasons for this may be a sense of curiosity, the goal of being more successful, and/or learning by trial-and-error. Hui (2023) stated in their findings that some students with high scores submitted the exams earlier with the aim of having additional time in case help was needed. Again, in the current study, students in the first group started their formative exams earlier than those in the other groups. This may have also indicated that their aim was to identify any misconceptions they may have had and correct those misconceptions within a sufficient amount of time required for the summative assessment. Students in Cluster 1 were also found to be the ones who spent the most amount of time on exam trials in terms of total time spent. Thus, as a result of the high number of trials, it was expected that the total time would increase as well. However, this could also be proof of their effort by considering their correct answers in subsequent attempts. As a result, students in this profile included those who were found to have the highest number of exam attempts. The fact that these students took the exams a considerable amount of time before the summative exam and regularly took the formative exams enabled them to increase their scores throughout this process. According to Hui (2023), trials that do not lead to progress in scoring can be described as trial-and-error, and trials that increase scores over time can be described as effortful improvement. Similarly, students with low scores show more random guessing behavior than those with high scores (Stenlund et al., 2017). In this respect, we can assume that students

in Cluster 1 exhibited some form of trial-and-error or random guessing behavior as part of their first trials and then likely worked to show effortful improvement progressively.

Next, Cluster 2 consisted of students who received the highest scores among all the clusters. In the formative exam, their first-attempt scores were higher, yet their subsequent attempt scores were lower. Importantly, the number of trials was low, along with the number of interactions being lower than that of other students. Students who studied and desired to assess themselves prior to the final exam might have been part of this group. Furthermore, according to their behavior, they tended to wait until the last day, just before the final exam. Interestingly, this pattern resembled a similar pattern from Stadler et al. (2020), which demonstrated that higher-ability students spent more time in the problem-solving process but interacted less than others. The fact that these students earned high scores, more recently tested themselves, and showed little interaction indicated having a higher level of ability. Also, we can conclude that students in this group were successful due to more likely knowing what they wanted and not needing to make further attempts and/or spend further time due to their earning high scores on the first attempt. This pattern was also similar to the result of Hui (2023) in which students with high scores did not make any subsequent attempts and ultimately stopped making attempts. Additionally, Hui (2023) explained that these students did not benefit from additional time and, as a result, should not make further attempts during this period due to the exam being too easy for them and, therefore, not needing to carry out additional work. The fact that students in Cluster 2 earned high scores on the formative exam on their first attempt and that their scores decreased in the next application followed a similar pattern. On the other hand, Yang et al. (2022) stated that students who frequently participate in formative assessment following the lesson receive higher scores from the summative exam than those who do not. A similar pattern was also observed in the current study when students from Cluster 2 received high scores for both the formative and summative exams.

In Cluster 3, students who had limited interaction were found to also score lower on formative assessments. Additionally, several students in this subgroup did not complete the final exam. Hui (2023) claims that some students may struggle to understand course material, which causes them to lose interest. This could be what caused the students to disengage as well as preventing them from completing the final exam. Importantly, a risk of dropping out of this course was present for students from this cluster. In this regard, it is important to intervene and assist students who fail to participate and/or receive low scores in their formative assessment, and this group of students should be considered at-risk for drop-out students.

Conclusion, Limitations, and Future Research

In the current study, the formative assessment test-taking behavior of students was analyzed along with investigating the student profiles. Clustering was applied to the metrics collected from formative assignment interactions, and then these were compared with the metrics in terms of the differing student profiles. Importantly, meaningful differences were found between students' formative assessment test-taking behaviors and summative test scores. Additionally, the research outcomes can enrich the literature by showcasing learners' interactions with the unique formative assessment design of the study and by discussing parallels with distinct test-taking profiles evident within only a limited body of work. Therefore, researchers and practitioners should be able to recognize the profile of Cluster 1 as a behavioral model that necessitates instituting precautions with high-risk assessments or as a trial-and-error behavioral model where interventions within the assessment design have the potential to enhance learning performance. Conversely, the profile of Cluster 2 can be appraised as students' being self-directed learners capable of overseeing their own learning progress. Thus, to incentivize these learners to attain a higher level of performance, assessment designs featuring progressively more challenging questions for each attempt can be implemented. On the other hand, in the literature, the Cluster 3 profile can be assessed as learners at risk of dropping out and awaiting solutions. As a result, to mitigate drop-out occurrences and increase engagement within the learning process, interventions such as support for

countering demotivation, provision of diverse tasks, and adaptation of assessment and content should be employed among students in this profile.

One of the limitations of the current study, in regard to generalizability, was the small sample size. Features of clusters should be compared with research results from large-scale samples, along with also confirming the comparison of summative assessment performance among clusters. Another limitation was related to the features used in clustering. In particular, by deepening the time metrics (response time for each question), clusters with divergent test-taking profiles can be obtained within the formative assessment. In the current study, students were encouraged through the formative test to increase their efforts toward learning. This may be due to the formative assessment being structured in a way that creates equivalent exams for students regarding each subject it covers. However, although it cannot be guaranteed that students take exams of equal difficulty in each attempt, validity and reliability concerns of the formative tests can be mitigated due to the random selection of questions.

Finally, in future studies, students within at-risk groups should be identified, and appropriate interventions should be applied to assist them. As a result, students can behave similarly to the successful students found in other clusters, and the contribution of these interventions can lead to further investigations. Additionally, more advanced metrics can be revealed for formative assessments held in the Moodle learning management system. Thus, due to these metrics, more detailed information can be obtained, especially in regard to the student test-taking strategies found in Clusters 1 and 2.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: The data usage in this study was approved by the Ankara University Ethics Committee, Social Science Sub Committee (Document number:160, dated 17.05.2021).

References

Blondeel, E., Everaert, P., & Opdecam, E. (2023). Does practice make perfect? The effect of online formative assessments on students' self-efficacy and test anxiety. *The British Accounting Review* (101189). https://doi.org/10.1016/j.bar.2023.101189

Cassady, J. C., & Gridley, B. E. (2005). The Effects of Online Formative and Summative Assessment on Test Anxiety and Performance. *The Journal of Technology, Learning and Assessment, 4*(1). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1648

Crisp, V., & Ward, C. (2008). The development of a formative scenario-based computer assisted assessment tool in psychology for teachers: The PePCAA project. *Computers & Education, 50*(4), 1509-1526. https://doi.org/10.1016/j.compedu.2007.02.004

Gašević, D., Greiff, S., & Shaffer, D. W. (2022). Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond. *Computers in Human Behavior, 134*(2022), 107304. https://doi.org/10.1016/j.chb.2022.107304

Guo, H. (2021). How Did Students Engage with a Remote Educational Assessment? A Case Study. *Educational Measurement: Issues and Practice, 41*(3), 58-68. https://doi.org/10.1111/emip.12476

Guo, H., & Ercikan, K. (2021). Comparing Test-Taking Behaviors of English Language Learners (ELLs) to Non-ELL Students: Use of Response Time in Measurement Comparability Research. *ETS Research Report Series, 2021*(1), 1-15. https://doi.org/10.1002/ets2.12340

Han, S., & Kang, H.-A. (2021). Sequential Monitoring of Aberrant Test-Taking Behaviors Based on Response Times. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim, *Quantitative Psychology* Cham. https://doi.org/10.1007/978-3-030-74772-5_7

Harlen, W., & James, M. (1997). Assessment and Learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*(3), 365-379. https://doi.org/10.1080/0969594970040304

Hui, B. (2023). *Are They Learning or Guessing? Investigating Trial-and-Error Behavior with Limited Test Attempts*. Paper presented at the LAK23: 13th International Learning Analytics and Knowledge Conference, Arlington, TX, USA. https://doi.org/10.1145/3576050.3576068

Lee, Y. H., & Haberman, S. J. (2016). Investigating Test-Taking Behaviors Using Timing and Process Data. *International Journal of Testing, 16*(3), 240-267. https://doi.org/10.1080/15305058.2015.1085385

Liao, M., Patton, J., Yan, R., & Jiao, H. (2021). Mining Process Data to Detect Aberrant Test Takers. *Measurement: Interdisciplinary Research and Perspectives, 19*(2), 93-105. https://doi.org/10.1080/15366367.2020.1827203

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation, 26*(5-6), 275-301. https://doi.org/10.1080/13803611.2021.1963940

Man, K., Harring, J. R., Ouyang, Y., & Thomas, S. L. (2018). Response Time Based Nonparametric Kullback-Leibler Divergence Measure for Detecting Aberrant Test-Taking Behavior. *International Journal of Testing, 18*(2), 155-177. https://doi.org/10.1080/15305058.2018.1429446

Palmen, L. N., Vorstenbosch, M. A. T. M., Tanck, E., & Kooloos, J. G. M. (2015). What is more effective: a daily or a weekly formative test? *Perspectives on Medical Education, 4*(2), 73-78. https://doi.org/10.1007/s40037-015-0178-8

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education, 9*(1), 10. https://doi.org/10.1186/s40536-021-00104-6

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction, 60*(2019), 154-165. https://doi.org/10.1016/j.learninstruc.2018.01.004

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Computers in Human Behavior, 60*(2016), 333–341. https://doi.org/10.1016/j.chb.2016.02.074

Sarac, M., & Loken, E. (2023). Examining patterns of omitted responses in a large-scale English language proficiency test. *International Journal of Testing, 23*(1), 56-72. https://doi.org/10.1080/15305058.2022.2070756

Shin, J., Chen, F., Lu, C., & Bulut, O. (2022). Analyzing students' performance in computerized formative assessments to optimize teachers' test administration decisions using deep learning frameworks. *Journal of Computers in Education*, *9*(1), 71-91. https://doi.org/10.1007/s40692-021-00196-7

Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *Trames. Journal of the Humanities and Social Sciences, 17(*67/62), 433-448. https://doi.org/10.3176/tr.2013.4.08

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31*(100335), 1-22. https://doi.org/10.1016/j.edurev.2020.100335

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior, 111*(2020), 106442. https://doi.org/10.1016/j.chb.2020.106442

Stenlund, T., Eklof, H., & Lyren, P. E. (2017). Group differences in test-taking behaviour: an example from a high-stakes testing program. *Assessment in Education-Principles Policy & Practice, 24*(1), 4-20. https://doi.org/10.1080/0969594x.2016.1142935

Stenlund, T., Lyren, P. E., & Eklof, H. (2018). The successful test taker: exploring test-taking behavior profiles through cluster analysis. *European Journal of Psychology of Education, 33*(2), 403-417. https://doi.org/10.1007/s10212-017-0332-2

Tempelaar, D., Rienties, B., Mittelmeier, J., & Nguyen, Q. (2018). Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior, 78*(2018), 408-420. https://doi.org/10.1016/j.chb.2017.08.010

Toetenel, L., & Rienties, B. (2016). Learning Design – creative design to visualise learning activities. *Open Learning: The Journal of Open, Distance and e-Learning, 31*(3), 233-244. https://doi.org/10.1080/02680513.2016.1213626

Vonderwell, S., Liang, X., & Alderman, K. (2007). Asynchronous Discussions and Assessment in Online Learning. *Journal of Research on Technology in Education, 39*(3), 309-328. https://doi.org/10.1080/15391523.2007.10782485

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education, 19*(2), 95-114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2006). An Application of Item Response Time: The Effort-Moderated IRT Model. *Journal of Educational Measurement, 43*(1), 19-38. https://doi.org/https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., Ma, L., Cronin, J., & Theaker, R. A. (2013). Student test-taking effort and the assessment of student growth in evaluating teacher effectiveness. *Annual conference of the American Educational Research Association*, San Francisco, CA,

Wise, S. L., & Gao, L. (2017). A General Approach to Measuring Test-Taking Effort on Computer-Based Tests. *Applied Measurement in Education, 30*(4), 343-354. https://doi.org/10.1080/08957347.2017.1353992

Wolsey, T. (2008). Efficacy of Instructor Feedback on Written Work in an Online Program. *EdMedia + Innovate Learning Online 2022, 7*(2), 311-329.

Xiong, Y., & Suen, H. K. (2018). Assessment approaches in massive open online courses: Possibilities, challenges, and future directions. *International Review of Education, 64*(2), 241-263. https://doi.org/10.1007/s11159-018-9710-5

Yang, A. C. M., Chen, I. Y. L., Flanagan, B., & Ogata, H. (2022). How students' self-assessment behavior affects their online learning performance. *Computers and Education: Artificial Intelligence, 3*(2022), 100058. https://doi.org/10.1016/j.caeai.2022.100058

Yildirim-Erbasli, S. N., & Bulut, O. (2020). The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. *Educational Research and Evaluation, 26*(7-8), 368-386. https://doi.org/10.1080/13803611.2021.1977152

Yildirim-Erbasli, S. N., & Bulut, O. (2022). Designing Predictive Models for Early Prediction of Students' Test-taking Engagement in Computerized Formative Assessments. *Journal Of Applied Testing Technology, 22*(2), 1-14. Retrieved from https://jattjournal.net/index.php/atp/article/view/167548