Journal of Data Science and Modeling, Vol. 3, No. 2, 169-199, June 2025

Research Manuscript

BRAF: A Behavioral Response Analysis Framework for Student Patterns in the Era of Digital Assessment

Seyede Fatemeh Noorani *[1], Fatemeh Orooji[2]
AmirHushangTajfar[1]

[1]Department of Information and Computer Engineering, Faculty of Technology and Engineering, Payame Noor University (PNU), P.O. Box 19395-4697, Tehran, Iran

[2]Department of Educational Sciences and Learning Technologies, Mehralborz University, Tehran, Iran

Recieved: 13/06/2026 Accepted: 14/09/2026

Abstract: In the era of digital assessment, learners' behavioral data offer new opportunities for analyzing response patterns. This study presents the Behavioral Response Analysis Framework (BRAF), integrating behavioral data and machine learning to analyze responses in a randomized sequential online test. Data from 282 students completing a six-item test were analyzed using t-tests, Pearson correlation, Isolation Forest, one-way ANOVA, and Random Forest.

Isolation Forest identified 29 students (10.3%) with anomalous response-position patterns, whose mean score was significantly lower than that of other students (3.76 vs. 4.70; $p < 0.001$). No significant difference was found between median-based PositionBias groups. Overall, 45.7% of students changed their answers at least once, with answer-changing associated with a 0.61-point lower mean score ($p < 0.001$). Random Forest assigned greater importance to TotalChanges than PositionBias (53.3% vs. 46.7%), although predictive performance was limited.

The findings suggest that multivariate analysis may improve screening for unusual response-position patterns. Answer-changing was associated with lower performance, but causality cannot be established. BRAF is therefore best viewed as a screening framework for targeted human review rather than a diagnostic tool.

Keywords: Behavioral Response Analysis, Positional Bias, Response Change, Isolation Forest, Randomized Sequential Online Test, Learning Analytics

Mathematics Subject Classification (2010): 62H30, 68T01.

*Corresponding Author: sf.noorani@pnu.ac.ir

1. Introduction

With the expansion of educational technologies and the development of e-assessment systems, behavioral data have gained importance in the era of electronic testing, and a large volume of learners' behavioral data is generated and stored during the assessment process Chen et al. (2024). Unlike traditional tests that primarily focus on students' final answers, digital environments enable the recording of behavioral details, including response time, question review patterns, answer changes, and the manner of interaction with the test. Analysis of these data can provide valuable information about students' cognitive processes and test-taking strategies Xu et al. (2024); Ulitzsch et al. (2024); Lindner and Greiff (2023); He and Cui (2025).

In this study, the Behavioral Response Analysis Framework (BRAF) is proposed, which combines behavioral data with machine learning algorithms to analyze student response patterns in a randomized sequential online test.

The importance of analyzing behavioral data in educational assessment is such that leading psychometric associations have deemed response process analysis essential for interpreting the constructs measured in tests He and Cui (2025); Jiao et al. (2018); Embretson (2023); Ercikan and Pellegrino (2017).

In recent years, machine learning and educational data mining have gained attention as powerful tools for uncovering hidden patterns in educational data Chen et al. (2026); Guo et al. (2024); Noorani et al. (2024). Recent research has also demonstrated the value of fine-grained process data in digital assessment, including response-related events, completion time, and response-review behavior, for investigating learners' cognitive and behavioral processes Arslan et al. (2026). Specifically, algorithms such as Isolation Forest are capable of identifying anomalous behaviors without requiring predefined thresholds Azizi et al. (2025); Maniyan et al. (2024). Recent research has demonstrated that this algorithm can operate with high accuracy in various domains, such as detecting anomalies in web traffic and environmental sensor data. This capability can play an important role in studying students' behavioral biases and detecting patterns that are not observable with traditional statistical methods.

Positional bias and answer change are two core phenomena that directly affect the validity and reliability of score interpretations in multiple-choice tests. Recent research has shown that option positioning can increase cognitive load and lead to differential performance across different student subgroups Li and Gao (2025). On the other hand, answer changing during a test has been recognized as an important cognitive and behavioral strategy, such that the design of computerized adaptive tests (CAT) is required to provide the possibility of review and answer change without reducing statistical efficiency Wang et al. (2019). However, despite these theoretical and technological advances, empirical studies that simultaneously and interactively analyze these two phenomena in the context of a randomized sequential online test remain limited.

Positional bias refers to a situation in which a student, regardless of item content, tends to select a specific option position (e.g., always the second). This phenomenon causes some students' responses to be influenced by the physical location of the option rather than its content. The main problem is that traditional methods for detecting this bias—such as the t-test based on the standard deviation of selected positions—are linear and uni-dimensional, and thus unable to accurately identify students with consistent positional patterns. This methodological weakness reveals the necessity of employing novel machine learning-based approaches.

Answer change is another phenomenon prevalent in electronic tests. Students typically have the ability to review and revise their answers. However, research on whether this behavior is beneficial or harmful has yielded inconsistent results. Some studies consider answer change useful and associated with improved performance, while others view it as a sign of doubt and anxiety. This inconsistency in findings is particularly evident in the context of a randomized sequential online test—where returning to previous questions is not allowed and option orders are randomly shuffled—underscoring the need for further investigation.

Despite the value of previous research, certain dimensions related to positional bias and answer change appear to have received less attention. The present study was designed to address these dimensions.

First, the main focus of studies on positional bias has been at the item level; in other words, most research has examined how the position of the correct option affects item difficulty and discrimination. However, detecting positional bias at the student level—i.e., identifying students who consistently select a particular option position—has received less attention. Thus, while we know that an item with the correct option in the middle position is easier, we do not know which students consistently select a particular position or how this behavior affects their scores.

Second, the methods employed to detect positional bias have been predominantly based on traditional statistics (e.g., t-tests). Due to their linear nature, these methods are incapable of identifying complex, multidimensional patterns of student behavior. In contrast, advanced anomaly detection algorithms such as Isolation Forest, which have proven their effectiveness in various domains, can identify abnormal behaviors without requiring manual threshold setting. However, the application of this algorithm to positional bias has been overlooked.

Third, previous research has typically examined positional bias and answer change separately, and their combined association with student performance has received less attention. This is despite the possibility that students exhibiting positional bias may also show different answer-changing patterns, and examining both behavioral characteristics together may provide a more comprehensive understanding of student response behavior.

Fourth, most studies have been conducted in the context of paper-based or simple computer-based tests. However, the characteristics of a randomized sequential online test, such as the inability to return to previous questions and the random shuffling of option orders, create a different structure that may moderate the impact of positional bias and answer change. Therefore, investigating these phenomena in such a context appears essential.

Examining answer change and positional bias is important because these behaviors can affect the validity of test score interpretations. Process data collected in electronic tests enable the identification of these behaviors Ulitzsch et al. (2024). Studies have shown that positional bias, even in high-stakes tests, has small but systematic effects on responses Lions et al. (2023) and can make items easier and less discriminating Ercikan et al. (2023). However, detecting this phenomenon at the student level using traditional methods is difficult and requires novel machine learning approaches Ercikan et al. (2023).

The BRAF framework in this study attempts to fill these gaps by leveraging behavioral data and machine learning algorithms. This study makes three primary contributions: (1) introducing a student-level anomaly detection approach to positional bias using Isolation Forest; (2) providing empirical evidence on the combined association of positional bias and answer change with student performance in a randomized sequential online test; and (3) offering a practical framework (BRAF) that can be integrated into existing e-assessment systems for behavioral screening and targeted human review rather than automated identification of at-risk students.

Based on the above, the main objectives of this study are:

To investigate the relationship between positional bias and student performance using both statistical and machine learning approaches; To examine the effect of answer change on student performance in a randomized sequential online test; To analyze the simultaneous and interactive effects of positional bias and answer change on student performance; and To determine the relative importance of each of these two factors in distinguishing normal students from those exhibiting abnormal behavior.

The remainder of this paper is organized as follows. First, the research methodology is presented, including the study population, variables, feature engineering process, and machine learning models (Isolation Forest, Random Forest). Next, the findings are reported at three levels (positional bias, answer change, and their combination). Finally, the results are discussed and interpreted, followed by practical implications, limitations, and suggestions for future research.

2. Literature Review

2.1 Response Bias and Validity in Educational Assessment

Response bias refers to systematic patterns in respondents' answers that are not fully attributable to the content being assessed and may therefore affect the interpretation of test scores. Cronbach and Meehl Cronbach and Meehl (1955) emphasized the importance of construct validity in interpreting psychological test scores, while Messick Messick (1995) highlighted the role of construct-irrelevant influences in the interpretation and use of assessment scores. In this context, systematic response-position preferences may represent a potential source of construct-irrelevant variation when they influence responses independently of the knowledge or ability targeted by the assessment. Accordingly, examining response behavior alongside final scores can provide complementary evidence for understanding how students interact with assessment items.

2.2 Positional Bias in Multiple-Choice Questions

Research in educational assessment has demonstrated that well-designed multiple-choice questions can be effective tools for assessing higher-order learning. However, one persistent challenge is positional bias, defined as an individual's systematic tendency to select a specific option position (e.g., always a selected choice) regardless of item content Brady (2005). This phenomenon indicates that some students' responses are influenced by the physical location of the option rather than its content.

The initial research in this area began with the question of whether the physical position of options affects examinees' responses and scores. In this regard, the authors of Blunch (1984) showed that simply rotating option order does not eliminate this bias, and its effect can be larger than sampling error.

In this direction, the study Attali and Bar-Hillel (2003) created a major breakthrough in the field. They demonstrated that both test designers and examinees have a systematic tendency toward middle positions. In a four-option test, approximately 55% of incorrect responses were located in the two central positions, B and C. The psychometric implication of this finding was significant: items with the correct answer in the middle position are easier and less discriminating.

However, the authors of Sonnleitner et al. (2016) moderated these findings through a study on national tests. They showed that the effect of correct option position exists but is small, and for older students and in high-stakes tests, positional effects are largely negligible.

The next major advance in this field was made by researchers in Lions Maitre et al. (2021). They demonstrated that attractive distractors (those very similar to the correct answer) are typically placed in middle positions B and C, while weak distractors are placed in the last position. This finding provides a new explanation for the "middle bias" phenomenon: the probability that a student selects a middle option occurs because attractive distractors are located in those positions, not because students inherently prefer middle options. Subsequently, Lions et al. Lions et al. (2022) showed that correct option position has small but systematic and replicable effects on responses. The researchers in Fagley (1987) also emphasized that the position of distractors affects item difficulty even when the correct option has no positional sensitivity.

In recent years, the study Maguya (2022) demonstrated that positional effects may vary depending on the test context and conditions, and may affect different test-takers differently.

The investigation of positional bias has extended beyond the human domain. For example, Jiao et al. (2026) showed that large language models also exhibit positional bias when answering multiple-choice questions. Unlike previous methods that focused on reducing this bias, they demonstrated that certain option permutations can actually improve model performance. This finding suggests that positional bias is a pervasive phenomenon, but its behavior in artificial systems differs from that in humans.

Reviewing previous research, it becomes evident that the main focus of studies on positional bias has been at the item level. In other words, most research has examined how the position of the correct option affects item difficulty and discrimination, or where attractive distractors are typically placed. However, detecting positional bias at the student level—i.e., identifying students who consistently select a particular option position—appears to have received less attention. While we know that "an item with the correct answer in the middle position is easier" or "attractive distractors are located in middle positions," the question of "which students consistently select a particular position and how does this behavior affect their scores?" remains largely unanswered.

It is important to distinguish positional bias from other commonly discussed response styles, such as extreme response style and midpoint response style. Extreme response style refers to a tendency to select the endpoints of a rating scale, whereas midpoint response style refers to a tendency to select the middle category, particularly in rating-scale instruments. Positional bias in the present study has a different operational meaning: it refers to a systematic tendency to select a particular physical position of the answer options in multiple-choice questions, irrespective of the semantic content of the options. Thus, the construct examined in this study concerns the spatial position of response alternatives rather than a preference for extreme or middle response categories.

2.3 Answer Change

Answer change refers to a situation in which a student, after selecting an option, changes their mind and selects a different option. This behavior may occur once or multiple times. In computer-based tests, unlike paper-based tests, precise recording and analysis of these changes is possible Wang et al. (2019). More broadly, recent learning analytics research has demonstrated that process data from digital assessments can be used to characterize students' test-taking behaviors beyond final response accuracy. For example, Guo et al. Guo et al. (2024) used machine-learning methods to derive process profiles from students' response, timing, and tool-use sequences, illustrating the potential of assessment process data for contextualizing student performance.

For years, the question has been whether changing answers in multiple-choice tests benefits students. In a classic study, Bauer et al. demonstrated that answer changes in paper-based tests predominantly occur from incorrect to correct Bauer et al. (2007). Merry et al. reported a similar finding: on average, students changed 2.8 times from incorrect to correct, whereas the reverse change (correct to incorrect) occurred only 1.0 times Merry et al. (2021). In any case, the effect of answer change is not uniform and also depends on whether the student is weak or strong Ferguson et al. (2002).

However, there is a fundamental problem: all of these studies were conducted in paper-based test contexts, where students can return to previous questions and change their answers. The transition to computer-based testing. On the other hand, researchers have sought the technical feasibility of designing tests that allow students to change their answers. Wang et al. Wang et al. (2019) proposed an approach demonstrating that response revision can be implemented in computerized adaptive tests (CAT) without reducing statistical efficiency. However, this study did not answer the question of "whether answer change in such tests benefits or harms the student."

2.4 Research Gap and Novelty

Previous studies have extensively examined positional effects in multiple-choice assessment, primarily at the item level, focusing on how the position of the correct option or distractors affects item difficulty, discrimination, and response selection Lions Maitre et al. (2021); Lions et al. (2022); Fagley (1987); Maguya (2022). In contrast, comparatively less attention has been given to identifying systematic response-position patterns at the student level across multiple items. Similarly, answer changing has been investigated in paper-based, randomized sequential online testing contexts Wang et al. (2019); Bauer et al. (2007); Merry et al. (2021); Ferguson et al. (2002), but it has generally been examined separately from positional behavior. Thus, the joint analysis of student-level positional patterns and answer-changing behavior in a sequential randomized digital assessment remains insufficiently explored.

To address these gaps, BRAF integrates student-level response-position analysis, unsupervised anomaly detection using Isolation Forest, and answer-change analysis within a single behavioral framework. Although machine-learning methods, including Isolation Forest, have been used for anomaly detection in educational data Azizi et al. (2025); Maniyan et al. (2024), the contribution of BRAF is not the introduction of a new algorithm; rather, it lies in applying anomaly detection to student-level response-position patterns and integrating this analysis with answer-change behavior and conventional statistical analysis. From a practical perspective, BRAF can be implemented as a screening layer in digital assessment systems to convert routinely collected process data into interpretable behavioral indicators and flag students who may benefit from further diagnostic review or targeted educational support. These indicators are intended as screening signals rather than definitive diagnoses of ability or learning difficulty.

3. Methodology

The aim of this study is to analyze student behavior in a randomized sequential online test at three levels: positional bias, answer change, and the combined examination of these two behavioral characteristics. To achieve this aim, the Behavioral Response Analysis Framework (BRAF) was designed and implemented. This framework was designed and executed in five stages. In the first stage, data from 282 students who participated in an online test with six questions were collected. In the second stage, two main behavioral indicators—namely, the positional bias indicator and the answer change indicator—were extracted separately for each student from the raw data. In the third stage, statistical analyses were performed, including independent t-tests to compare mean scores between groups, Pearson correlation to examine the relationship between the number of changes and final score, and ANOVA to compare the four combination groups. In the fourth stage, machine learning algorithms were employed: Isolation Forest for detecting behavioral anomalies at the student level and Random Forest for predicting student pass/fail status. In the fifth stage, the findings from the statistical and machine learning approaches were compared and interpreted. Figure 1 presents a schematic summary of these five stages. Each of these stages is described in detail below.

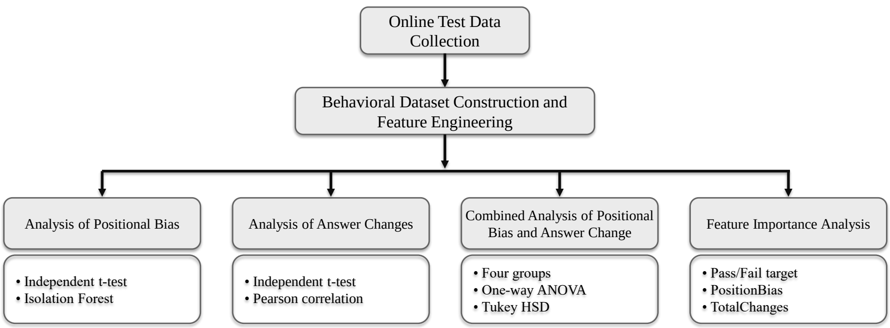


Figure 1: Overview of the proposed Behavioral Response Analysis Framework (BRAF).

3.1 Online Test Data Collection

The present study was conducted on data from 282 undergraduate computer engineering students enrolled in the "Theory of Languages and Machines" and "Discrete Mathematics" courses at Payame Noor University during the second semester of 1404-1405 (2025-2026 academic year), who participated in a randomized sequential online test. The test consisted of six questions from six different content groups, labeled A through F. Each group contained two versions of the question (Version 1 and Version 2), from which the system randomly selected one for each student. The questions were of equal difficulty level, and the total test score was 6.

The test was administered in a computerized, sequential manner, meaning that after answering each question, the student could not return to previous questions, and only one question was displayed at any given time. The order of options and the sequence of question presentation varied across students. Table 1 presents the demographic and statistical characteristics of the participants.

Table 1: Statistical characteristics of the study population

Indicator	Value
Number of students	282
Mean score	4.60 (out of 6)
Standard deviation of scores	1.42
Pass rate (score $\geq$ 4)	76.2%

3.2 Behavioral Dataset Construction and Feature Engineering

To conduct the research analyses, data were collected from three different sources within the test. The first source was each student's test results, indicating how many questions the student answered correctly, incorrectly, and left unanswered (blank), as well as their total score out of 6 (Figure 2). The second source was the question details. In the test administration system, clicking the "View" button allowed access to each student's test details (Figure 2). This section revealed: the order of questions for each student, the order of options, which question from each group was shown to the student, as well as the correct answer and the option selected by the student.

The third source was the test log file (Figure 2). This log recorded the test entry time, the time each answer was recorded, and the number of times different answers were selected for each question. These three files were used to construct the dataset for this study (student identifiers have been masked to protect privacy).

The features extracted from the three sources above are shown in Table 2.

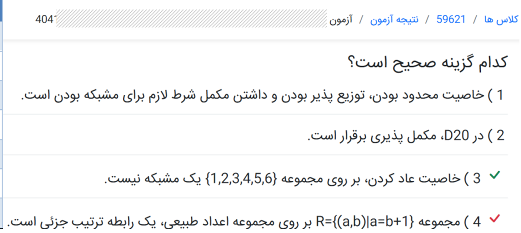
Table 2: Data collected from the three test sources

Category	Fields	Description
Basic Information	StudentID, Name, Fname, Course	Includes student ID, first name, last name, and course name
Performance	BlankCounts, MistakeCounts, CorrectCounts, Mark	Number of unanswered questions, number of incorrect answers, number of correct answers, and final score
Student's Questions	q1, q2, q3, q4, q5, q6	For questions 1 through 6: which group (A-F) and which of the two questions from each group
Student's Answers	Ans1, Ans2, Ans3, Ans4, Ans5, Ans6	Student's answer to questions 1 through 6. 1 = correct, 0 = incorrect, 3 = blank
Correct Option Position	CPq1, CPq2, CPq3, CPq4, CPq5, CPq6	Position of the correct option for questions 1 through 6 for each student (1 to 4)
Selected Option Position	SPq1, SPq2, SPq3, SPq4, SPq5, SPq6	The option selected by the student for questions 1 through 6 (1 to 4)

To improve model predictive power and extract hidden behavioral patterns, the following features were extracted from the raw data at this stage:

نمره	صحیح	غلط	سفید	نام خانوادگی	نام	نام کاربری
2.00	2	4	0	[redacted]	سعیده	40*****38
4.00	4	2	0	[redacted]	نگین	40*****57
6.00	6	0	0	[redacted]	موحد	40*****51
5.00	5	1	0	[redacted]	مبینا	40*****85
3.00	3	3	0	[redacted]	محمد	40*****60
6.00	6	0	0	[redacted]	ایدا	40*****84
3.00	3	3	0	[redacted]	حسن	40*****64

(a) Student results table



(b) Question details view

فعالیت	نوع	IP	شماره دانشجو	تاریخ/زمان
match:26821-qid:16549-status:True-type:AutoSave	ثبت پاسخ	192.0.2.101	4*******6	۲۵/۰۳/۱۴۰۵ ۱۰:۳۲
match:26821-qid:16549-status:True-type:AutoSave	ثبت پاسخ	192.0.2.102	4*******6	۲۵/۰۳/۱۴۰۵ ۱۰:۳۲
match:26821-qid:16362-status:True-type:AutoSave	ثبت پاسخ	192.0.2.105	4*******7	۲۵/۰۳/۱۴۰۵ ۱۰:۳۲
match:26821-qid:16559-status:True-type:AutoSave	ثبت پاسخ	192.0.2.106	4*******5	۲۵/۰۳/۱۴۰۵ ۱۰:۳۲
Online Test Activity	ورود به فعالیت	192.0.2.107	9*******5	۲۵/۰۳/۱۴۰۵ ۱۰:۳۲
match:26821-qid:16559-status:True-type:AutoSave	ثبت پاسخ	192.0.2.108	4*******9	۲۵/۰۳/۱۴۰۵ ۱۰:۳۳
match:26821-qid:16576-status:True-type:AutoSave	ثبت پاسخ	192.0.2.109	4*******4	۲۵/۰۳/۱۴۰۵ ۱۰:۳۳
match:26821-qid:16399-status:True-type:AutoSave	ثبت پاسخ	192.0.2.110	4*******3	۲۵/۰۳/۱۴۰۵ ۱۰:۳۳

(c) Test log file

Figure 2: Three data sources used to construct the dataset for this study. *Note: For privacy and anonymization purposes, student identifiers, activity IDs (match), and question IDs (qid) have been partially masked, while all IP addresses have been replaced with simulated values. The displayed values are illustrative and do not correspond to identifiable students or actual system records.*

Positional Bias: To obtain a student-level indicator of positional response consistency, the standard deviation of the student's selected option positions across the six questions was calculated. Specifically, PositionBias was computed from the six selected-option-position variables (SPq1–SPq6) using the following formula:

$$PositionBias_i = \sqrt{\frac{1}{6}\sum_{j=1}^{6}(SPq_{ij} - \overline{SPq_i})^2}$$

Where:

- SPq_{ij} is the position of the option selected by student i on question j (a value between 1 and 4)

- $\overline{SPq_i}$ is the mean of the selected positions for student i across the six questions

A low value (close to zero) indicates greater positional consistency, whereas a high value indicates greater dispersion across option positions. Accordingly, lower values were operationally treated as an indicator of potential positional bias in the statistical analysis. This classification is used as a screening rule and should not be interpreted as a definitive identification of abnormal behavior.

Blank responses were coded as 0 in the dataset and retained in the calculation of the PositionBias indicator. Given their very small number (7 out of 1,692 student-question responses), these blank responses were considered unlikely to have a meaningful effect on the PositionBias calculations.

Importantly, PositionBias is a derived, single student-level feature rather than a direct representation of the six SPq variables. The six SPq1–SPq6 variables are first combined through the above formula to obtain one PositionBias value for each student. Because standard deviation captures dispersion only, PositionBias does not fully characterize all possible non-random response-position structures.

Answer Change Indicator: In this study, the total number of answer changes for each student was calculated by summing the number of changes across the six questions. This information was extracted programmatically from the third source, the test log file shown in Figure 2. Because the test was administered sequentially and students were not permitted to return to previous questions, all recorded answer changes occurred while the current question was displayed.

An important note: In this study's dataset, a value of 1 for the variable NumberofAnswerChangeq means that the question was answered once and no change occurred. Values greater than 1 indicate the number of answer changes. Therefore, the following formula was used to calculate the total number of changes across six questions for each student:

$$TotalChanges_i = \sum_{j=1}^{6} (NumberofAnswerChangeq_{ij} - 1)$$

Where:

- $NumberofAnswerChangeq_{ij}$ is the number of times student i changed their answer on question j

- A value of 1 = no change (i.e., the student answered only once and did not change)

- A value of 2 = one answer change (i.e., the student tried two different options)

Target Variable: To convert the prediction problem into a binary classification problem, a pass/fail variable (Pass) was defined. Students who scored 4 or higher (out of 6) were considered passing, and the remaining were failing:

$$Pass = \begin{cases} 1 & \text{if } CorrectCounts \geq 4 \\ 0 & \text{otherwise} \end{cases}$$

The reason for selecting a threshold of 4 is that a score of 4 out of 6 is equivalent to 66.7%, which is considered a passing grade in many educational systems. This threshold also creates a relatively balanced distribution between the two classes (76.2% passing, 23.8% failing).

3.3 Analysis of Positional Bias

At the first level of analysis, two complementary approaches were employed: a statistical approach based on the independent t-test and a machine learning approach based on the Isolation Forest algorithm.

Statistical Approach (t-test): First, the PositionBias indicator was calculated for each student. This indicator, derived from the standard deviation of the student's selected option positions across the six questions, reflects the dispersion of selected option positions. Students were then divided into two groups based on the sample median of this indicator: a lower-dispersion PositionBias group and a higher-dispersion PositionBias group. Students with PositionBias values below the median were classified into the lower-dispersion PositionBias group, whereas those with values at or above the median were classified into the higher-dispersion PositionBias group. This classification reflects differences in the dispersion of selected positions but does not imply confirmed presence or absence of positional bias. Finally, an independent t-test was conducted to compare the mean scores of the two groups. To determine statistical significance, the p-value threshold was set at 0.05, and a 95% confidence interval for the mean difference was calculated.

Machine Learning Approach (Isolation Forest): The Isolation Forest algorithm is an unsupervised method for anomaly detection introduced by Liu et al. (2008). Unlike traditional methods that attempt to model the "normal pattern" and then identify anomalies based on deviation from that pattern, Isolation Forest directly seeks to "isolate" anomalies Liu et al. (2008). The core idea of this algorithm is straightforward: anomalies are samples that can be isolated with fewer splits compared to other data points.

The algorithm randomly selects a feature and then a split value between the minimum and maximum of that feature, partitioning the data into two subsets. This process is repeated recursively until each sample is completely isolated. The path length (number of splits) required to isolate a sample serves as a measure of its "anomalousness." Samples that are isolated with fewer splits are considered more anomalous because their features differ significantly from other data points.

In this study, each student's selected option positions across the six questions (SPq1 to SPq6) were used as input features to the algorithm. These six features represent the student's response pattern throughout the test. A student with a normal pattern (diverse positions) will have a longer isolation path, whereas a student with positional bias (always selecting a repeated position) will have a shorter isolation path and be identified as an anomaly.

This use of the six SPq variables applies specifically to the Isolation Forest analysis. In the subsequent Random Forest analysis, the six SPq variables were not entered separately; instead, they were summarized into the single derived PositionBias feature described above.

To configure the Isolation Forest, the contamination parameter was set to 0.10 and the number of estimators to 100. The contamination value was selected as a theoretically plausible screening rate for rare behavioral patterns and was further examined through a sensitivity analysis using contamination values of 0.05, 0.10, and 0.15. Across these values, the identified anomalous groups consistently showed lower mean scores than the corresponding normal groups, supporting the stability of the observed association. The 0.10 setting was retained as the primary specification because it provided a moderate anomaly rate while avoiding the potential under-identification at 0.05 and over-identification at 0.15.

3.4 Analysis of Answer Change

In the testing environment, answer change can indicate doubt, uncertainty, or re-evaluation of a question. To investigate this phenomenon, two methods were used: the independent t-test and Pearson correlation.

Independent t-test: This method addresses whether there is a significant difference between the mean scores of students who changed their answers and those who did not. After calculating the TotalChanges indicator for each student, students were divided into two groups: the "no change" group (TotalChanges $= 0$) and the "change" group (TotalChanges ≥ 1). Finally, an independent t-test was performed to compare the mean scores of the two groups. The significance level (p-value) was set at 0.05, and a 95% confidence interval for the mean difference was calculated.

Pearson Correlation: The second method, Pearson correlation, addresses whether a linear relationship exists between the "number of answer changes" and the "final score," and, if a relationship exists, whether its direction is positive or negative. Pearson's correlation coefficient (r) was calculated at the student level between the total number of changes (TotalChanges) and the final score (CorrectCounts). A negative coefficient indicates an inverse relationship (the higher the number of changes, the lower the score). To determine significance, the p-value threshold was set at 0.05.

3.5 Combined Analysis of Positional Bias and Answer Change

This analysis examined whether PositionBias and answer-changing behavior, when considered jointly, were associated with differences in student performance. For this four-group analysis, PositionBias was classified using the sample median as the threshold. Students with PositionBias values below the median were assigned to the lower-dispersion PositionBias group, whereas students with PositionBias values at or above the median were assigned to the higher-dispersion PositionBias group. This median-based classification was used specifically for the four-group comparison and is distinct from the Isolation Forest procedure, which was applied separately to identify students with anomalous response-position patterns.

Students were divided into four groups based on the combination of the median-based PositionBias classification and answer-changing behavior, as shown in Table 3. The two dimensions used for this classification were response-position dispersion and answer-changing behavior.

Table 3: Categorization of participants by PositionBias dispersion group and answer-changing behavior

Group	PositionBias Group	Answer Change	Description
1	Higher-dispersion PositionBias	No	Students in the higher-dispersion PositionBias group who did not change their answers (reference group)
2	Higher-dispersion PositionBias	Yes	Students in the higher-dispersion PositionBias group who changed their answers at least once
3	Lower-dispersion PositionBias	No	Students in the lower-dispersion PositionBias group who did not change their answers
4	Lower-dispersion PositionBias	Yes	Students in the lower-dispersion PositionBias group who changed their answers at least once

The terms ‘lower-dispersion” and ‘higher-dispersion” refer only to the student’s PositionBias value relative to the sample median. They describe relative response-position dispersion and should not be interpreted as confirmed presence or absence of positional bias.

After forming these four groups, a one-way ANOVA* was conducted to compare the mean scores across the four groups. The purpose of this analysis was to

*Analysis of Variance

determine whether there was a statistically significant difference in mean performance among the four operational groups.

The null hypothesis (H_0) in the ANOVA is that all group means are equal. If the null hypothesis is rejected ($p < 0.05$), it can be concluded that at least two group means differ significantly from each other. The significance level was set at 0.05, and the F-statistic and corresponding p-value were reported.

The four-group analysis was interpreted as a combined group comparison rather than as a formal statistical test of an interaction effect between PositionBias and answer change. Therefore, differences among the four group means do not by themselves establish a statistical interaction between the two factors or a causal effect of either factor on student performance.

3.6 Implementation Tools

All preprocessing, feature engineering, modeling, and evaluation steps were performed using Python version 3.10 in the Jupyter Notebook environment, with the following libraries: Pandas version 2.0, NumPy version 1.24, Scikit-learn version 1.3, SciPy version 1.10, Matplotlib version 3.7, and Seaborn version 0.12.

4. Results

In this section, the outcomes of the Behavioral Response Analysis Framework (BRAF) are presented. The results are organized according to the three levels of analysis specified in the research objectives: positional bias, answer change, and the combined effect of both factors.

4.1 Analysis of Positional Bias

As previously discussed, in the statistical approach (t-test), the median value was used as the threshold. Selecting the median as the threshold is a standard method in data analysis that divides the data distribution into two equal parts and is robust to outliers Wilcox (2012). Accordingly, students whose PositionBias values were below the median (136 students) were classified into the lower-dispersion PositionBias group, whereas the remaining students (146 students) were classified into the higher-dispersion PositionBias group. An independent t-test was conducted to compare the mean scores of the two groups.

As shown in Table 4, the results indicated that the mean scores of the higher-dispersion PositionBias group (4.59) and the lower-dispersion PositionBias group (4.61) were nearly equal, and there was no significant difference between them (t = -0.131, p = 0.896).

Table 4: Results of the statistical approach (t-test) for positional bias

Group	N	Percentage	Mean Score	Standard Deviation
Higher-dispersion PositionBias group	146	51.8%	4.59	1.41
Lower-dispersion PositionBias group	136	48.2%	4.61	1.43
Difference		+0.02		
t-statistic		-0.131		
p-value		0.896		

The effect size was negligible (Cohen's d = -0.016), indicating that the observed difference in performance between the two groups was practically insignificant.

As the second analysis of positional response patterns, the Isolation Forest algorithm, which does not require manual threshold selection, was used. By considering the joint pattern of the six selected option positions for each student, the algorithm identified 29 students (10.3% of the total sample) as having anomalous response-position patterns (3). The mean score of these students (3.76) was lower than the mean score of the remaining students (4.70) and also lower than the overall sample mean (4.60). The 0.94-point difference was statistically significant and represents a notable difference in observed performance. However, these results should not be interpreted as definitive evidence that all 29 students exhibited positional bias. Rather, they indicate that these students showed response-position patterns that were unusual relative to the rest of the sample.

A sensitivity analysis was conducted to examine the effect of the contamination rate on the Isolation Forest results. As shown in Table 5, contamination rates of 0.05, 0.10, and 0.15 identified 15 (5.32%), 29 (10.28%), and 43 (15.25%) students as anomalies, respectively. In all three settings, the anomaly group had a significantly lower mean score than the corresponding normal group ($p < 0.01$). Although the magnitude of the difference varied across contamination rates, the direction of the association remained consistent.

Table 5: Sensitivity analysis of the Isolation Forest contamination parameter

Contamination	Anomaly N	Anomaly (%)	Normal Mean	Anomaly Mean	Difference	t	p
0.05	15	5.32	4.67	3.27	1.41	4.137	0.00080
0.10	29	10.28	4.70	3.76	0.94	3.601	0.00098
0.15	43	15.25	4.70	4.05	0.65	2.784	0.00732

In the statistical approach, due to the use of the median as the threshold and the forced 50-50 split, the lower-dispersion PositionBias group included many students whose response-position patterns may not necessarily have been unusual, leading to the failure to observe a significant difference. In contrast, the Isolation Forest algorithm, without relying on the standard-deviation threshold, identified

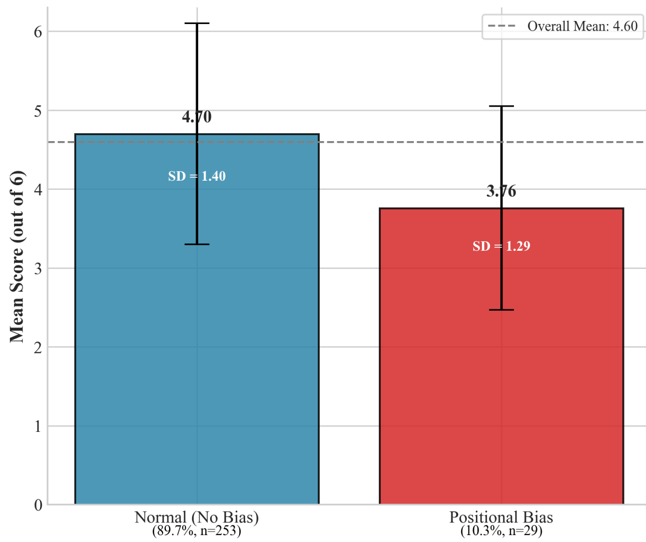


Figure 3: Comparison of mean scores between normal students (N=253, M=4.70, SD=1.40) and students identified with positional bias (N=29, M=3.76, SD=1.29) using the Isolation Forest algorithm. Error bars represent standard deviations.

29 students with comparatively unusual response-position patterns. This finding suggests that multivariate analysis of the complete sequence of selected positions may provide a more informative screening approach than a univariate dispersion-based rule. Nevertheless, the identified cases should be interpreted as anomalous response patterns rather than definitive evidence of positional bias or abnormal student behavior.

4.2 Analysis of Answer Change

Figure 4 presents the descriptive statistics of answer changes in the test. Out of 282 students, 129 (45.7%) changed their answers at least once, and 153 (54.3%) made no changes to their answers. The mean total number of changes per student was 1.26, and the mean number of questions in which answer changes occurred was 0.70. The maximum number of changes observed for a single question was 11. These statistics indicate that although nearly half of the students changed their answers at least once, these changes were typically concentrated on a limited number of questions (fewer than one question per student).

In the t-test, to examine whether answer change is associated with student performance, students were first divided into two groups: "no change" (153 students) and "change" (129 students). An independent t-test was then conducted to

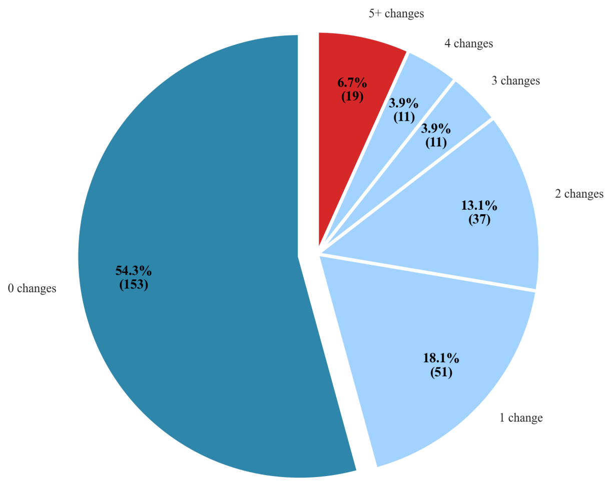


Figure 4: Distribution of answer changes among students (N=282). The majority of students (54.3%, n=153) made no changes to their answers, while 45.7% changed their answers at least once. Only 6.7% (n=19) changed their answers five or more times. The dashed line indicates the mean number of changes per student (1.26).

compare the mean scores of the two groups. The results showed that students who changed their answers achieved a lower mean score (4.27) compared to students who made no changes (4.88). This 0.61-point difference was statistically significant ($t = 3.81$, $p < 0.001$). In other words, answer change in this test was associated with decreased performance.

The effect size was small to moderate (Cohen's $d = 0.456$), suggesting a modest practical difference in performance between students who changed their answers and those who did not.

In the second step, Pearson's correlation coefficient was used between the total number of changes (TotalChanges) and the final score (CorrectCounts). The purpose of this analysis was to answer whether a linear relationship exists between the "number of answer changes" and the "final score." The results showed a significant negative correlation between these two variables ($r = -0.143$, $p = 0.017$). This means that as the number of answer changes increases, the student's final score decreases. Although this relationship is weak, it is statistically significant and confirms the t-test finding.

The correlation coefficient itself represents the effect size and indicates a small negative association between the number of answer changes and student performance.

4.3 Combined Analysis of Positional Bias and Answer Change

At this level, two aspects were examined: (1) the combined association of response-position dispersion and answer-changing behavior with student performance, and (2) the extent to which the four operational groups differed in their observed performance.

Figure 5 shows the mean score and standard deviation for each of the four groups. The results of the one-way ANOVA revealed a statistically significant difference among the four groups (F = 4.84, p = 0.003, η^2 = 0.050), indicating that group membership accounted for approximately 5% of the variance in student performance.

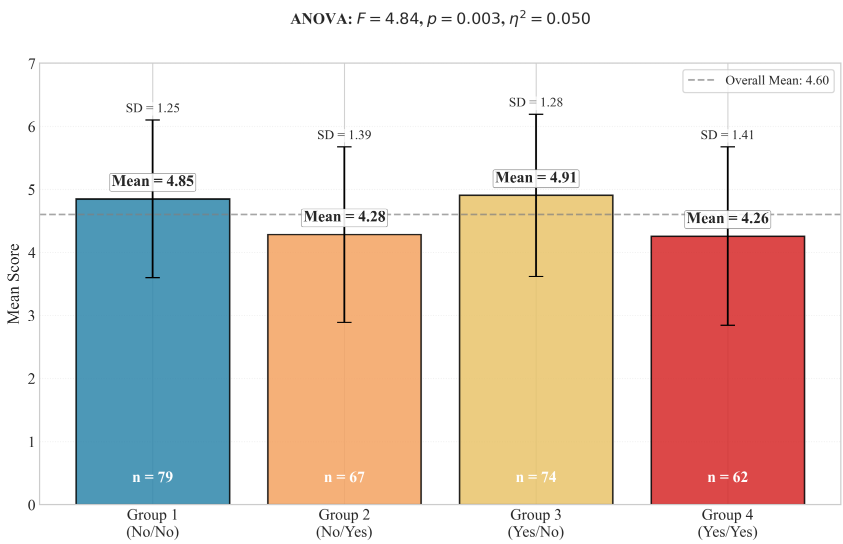


Figure 5: Comparison of mean scores across four student groups categorized by PositionBias dispersion and answer-changing behavior (N=282). Group 3 (lower-dispersion PositionBias, no change) achieved the highest mean score (4.91, SD=1.28, n=74), whereas Group 4 (lower-dispersion PositionBias, with change) achieved the lowest mean score (4.26, SD=1.41, n=62). Group 1 (higher-dispersion PositionBias, no change) had a mean score of 4.85 (SD=1.25, n=79), and Group 2 (higher-dispersion PositionBias, with change) had a mean score of 4.28 (SD=1.39, n=67). The error bars represent standard deviations, and the dashed line indicates the overall sample mean. The overall difference among the four groups was statistically significant (F = 4.84, p = 0.003, η^2 = 0.050).

As illustrated in Figure 5, the highest mean score was observed in Group 3 (lower-dispersion PositionBias, no change; M = 4.91, SD = 1.28, n = 74), whereas the lowest mean score was observed in Group 4 (lower-dispersion PositionBias, with change; M = 4.26, SD = 1.41, n = 62). Group 1 (higher-dispersion PositionBias, no change) also showed a relatively high mean score (M = 4.85, SD = 1.25, n = 79), while Group 2 (higher-dispersion PositionBias, with change) had a lower mean score (M = 4.28, SD = 1.39, n = 67).

The effect size was $\eta^2 = 0.050$, indicating that approximately 5% of the variance in student performance was associated with group membership.

To determine which specific groups differed significantly, Tukey's honestly significant difference (HSD) post-hoc test was conducted. The results showed significant differences between Group 1 and Group 4 (adjusted p = 0.046), between Group 2 and Group 3 (adjusted p = 0.030), and between Group 3 and Group 4 (adjusted p = 0.026). No statistically significant differences were observed for the remaining pairwise comparisons after adjustment for multiple comparisons.

Overall, the group comparisons suggest that answer-changing behavior was associated with lower performance in this sample, as both groups involving answer changes (Groups 2 and 4) had lower mean scores than the two groups without answer changes (Groups 1 and 3). However, because the lower- and higher-dispersion PositionBias groups were defined using a median-based classification, these results should be interpreted as associations between the operational behavioral categories and observed performance. They do not establish an independent effect of response-position dispersion or positional bias on student performance.

4.3.1 Feature Importance Analysis

As a supplementary analysis, a Random Forest classifier was employed to examine whether the two student-level behavioral indicators, PositionBias and TotalChanges, contained complementary information for classifying students according to their Pass/Fail status. Before presenting the results, it is important to distinguish this analysis from the Isolation Forest analysis described in the previous section. Isolation Forest was applied directly to the SPq1–SPq6 variables to identify students whose response-position patterns were unusual within the sample. In contrast, Random Forest was used here for a different purpose: to examine the classification of Pass/Fail status using the two derived behavioral indicators, PositionBias and TotalChanges. Thus, the two machine-learning analyses address different analytical objectives.

The Random Forest model was trained using 80% of the observations (225 students) and evaluated on the remaining 20% (57 students). The target variable was Pass/Fail, and PositionBias and TotalChanges were used as predictors. On the test set, the model achieved an accuracy of 0.579, balanced accuracy of 0.456, F1-score of 0.714, and ROC-AUC of 0.321. These results indicate limited classification performance on the held-out test data. Therefore, the Random Forest analysis is considered supplementary and exploratory, and its results should not be interpreted as evidence of stable or generalizable predictive performance.

4.4 Sensitivity Analysis for Students with High Numbers of Answer Changes

Four students had unusually high numbers of answer changes, exceeding 10 during the test. These cases were treated as high-change observations and were examined separately in a sensitivity analysis rather than being automatically excluded as statistical outliers. The mean score of these students was 2.25, which is lower than the overall mean (4.60). Table 6 shows the information for these students. In this table, Total Number of Changes indicates the total sum of all changes made to the questions. Question with Maximum Changes and Maximum Changes on One Question show which question had the highest number of changes and the exact count.

Table 6: Students with very high numbers of answer changes

Student ID	Total Number of Changes	Maximum Changes on One Question	Question with Maximum Changes
40******1	14	11	Question 5
40******4	15	5	Question 3
40******1	12	7	Question 1
40******0	12	5	Question 3

To examine whether the observed correlation was influenced by students with unusually high numbers of answer changes, a sensitivity analysis was conducted. Four students (1.42% of the sample) had more than 10 answer changes, with 14, 12, 12, and 15 changes, respectively. These cases were retained in the primary analysis. When all 282 students were included, the Pearson correlation between TotalChanges and final score was (r=-0.1428) ((p=0.0164)). After excluding the four students with more than 10 answer changes, the correlation remained negative and statistically significant ((N=278), (r=-0.1370), (p=0.0223)). The difference in the correlation coefficient was small ((Δ r=0.0057)), indicating that the observed negative association was relatively robust to the exclusion of these high-change cases. Accordingly, these observations were not considered sufficient grounds for excluding the cases from the primary analysis.

5. Discussion and Conclusion

This study examined student response behavior in a randomized sequential online test by analyzing positional response patterns and answer-changing behavior. The findings showed that the Isolation Forest algorithm identified 29 students (10.3%) with anomalous response-position patterns, whose mean score (3.76) was significantly lower than that of the remaining students (4.70). These students should not be interpreted as definitively having Positional Bias; rather, they were identified because their joint response-position patterns across the six questions were unusual relative to the rest of the sample. In addition, 45.7% of students changed their answers at least once, and students who changed their answers had mean scores that were 0.61 points lower than those of students who did not change their answers ($p < 0.001$). In the four-group comparison, the lower-dispersion PositionBias group with answer changes had the lowest observed mean score (4.26), whereas the lower-dispersion PositionBias group without answer changes had the highest observed mean score (4.91). Because the four groups were defined using the median-based PositionBias classification, these results describe associations between the operational behavioral categories and observed performance rather than evidence of causal or independent effects of Positional Bias.

These findings suggest that positional response patterns and answer-changing behavior provide complementary information about student behavior in digital assessments. The negative association between answer changing and performance is consistent with studies suggesting that answer revision is not uniformly beneficial Vispoel et al. (2005), but differs from studies reporting positive effects of answer revision Bauer et al. (2007); Ferguson et al. (2002). This difference may be related to the characteristics of the present assessment context. The test was randomized, and students could not return to previous questions. Under such conditions, answer changes may be associated with uncertainty or inefficient decision-making rather than productive review, although the present data do not allow this interpretation to be tested directly. Thus, the association between answer changing and performance should be interpreted in relation to the assessment context and individual differences rather than as universally beneficial or harmful. Because the study is observational, the findings do not establish that answer changing causes lower performance. Alternative explanations, including the possibility that lower-performing students may be more likely to change their answers, cannot be ruled out.

The findings are also broadly consistent with recent research emphasizing the value of process data for understanding student behavior in digital assessment. The study of Arslan et al. (2026) demonstrated that process data can provide insight into the relationship between response behavior and underlying cognitive processes, while Guo et al. (2024) emphasized the importance of contextualizing test performance using richer assessment data. The present study extends this perspective by jointly examining response-position patterns and answer-changing behavior within a single analytical framework.

From a practical perspective, BRAF can be implemented as a behavioral screening layer within a digital assessment system rather than as an automated diagnostic tool. A practical implementation can follow a three-step workflow. First, the assessment system records response events, including the selected option position for each item and all answer revisions. Second, after the assessment is completed, BRAF automatically calculates the behavioral indicators and applies the predefined screening procedures: the response-position sequence is evaluated using the Isolation Forest model, while the total number of answer changes is calculated from the recorded response events. Third, students identified by these procedures are flagged for human review rather than being automatically classified as having a learning difficulty or other problem.

The resulting report could provide the instructor or assessment specialist with a small set of interpretable indicators, including the student's final score, total number of answer changes, the sequence of selected option positions, and whether the response-position pattern was flagged as anomalous. For example, a student with a low score and a high number of answer changes could be referred for further review of content understanding or test-taking strategy. In contrast, a student flagged only because of an unusual response-position pattern, but who has satisfactory performance and no other concerning behavioral indicators, would not necessarily require intervention. Thus, the purpose of BRAF is to support targeted review and the use of additional evidence, rather than to make automatic decisions about students.

In this screening workflow, the term "at-risk" should be used cautiously. BRAF does not establish that a flagged student is academically at risk, and neither Positional Bias nor answer-changing behavior should be used as a stand-alone criterion for assigning such a label. Instead, these indicators can be used to trigger a predefined review process in which the instructor examines the student's performance and, where appropriate, considers additional diagnostic assessment, feedback, or instructional support. This approach makes the practical role of BRAF explicit: it is intended to prioritize cases for human review, not to replace professional judgment or make automated decisions about individual students.

The findings may also inform the design of digital assessment interfaces. However, these implications should be regarded as hypotheses for future research rather than recommendations supported by the present study. For example, future studies could investigate whether an assessment system that provides a clear indication that a response has been revised and remains subject to final submission, or that requires explicit confirmation before submitting a revised response, can reduce accidental or uncertain response changes. Such interface features may be theoretically relevant because they could help distinguish deliberate revision from accidental or uncertain response changes, but their effectiveness has not been evaluated in the present study. Accordingly, the current findings provide no empirical evidence that a final-confirmation mechanism, or any other specific interface modification, would improve performance, reduce unproductive answer changes, or improve decision quality. These possibilities should therefore be tested experimentally against appropriate control conditions before being considered evidence-based design recommendations.

The Random Forest feature-importance results require a separate interpretation from the Isolation Forest findings. In the Isolation Forest analysis, SPq1–SPq6 were entered as six separate input features so that the joint structure of the student's response-position sequence could be evaluated. In contrast, the Random Forest model used to examine pass/fail status was fitted using two student-level behavioral features: TotalChanges and PositionBias. PositionBias was calculated from SPq1–SPq6 using the standard-deviation-based formula defined in the methodology. Thus, the six SPq variables were used to construct the single PositionBias feature and were not entered separately into this Random Forest model.

The relative feature-importance values of 53.3% for TotalChanges and 46.7% for PositionBias therefore refer specifically to the two features entered into the Random Forest model. The 46.7% value should not be interpreted as evidence that PositionBias is a complete measure of Positional Bias or that it independently explains 46.7% of the variance in student performance. Rather, it indicates the relative contribution of the derived PositionBias feature to the fitted Random Forest model in this dataset. Although PositionBias captures only the dispersion of selected option positions, it can still contain information relevant to distinguishing pass/fail outcomes. This does not contradict the limitation that dispersion alone cannot characterize every possible non-random response-position structure.

Moreover, the feature-importance values should be interpreted as model-based descriptive results rather than as evidence of causal importance. The Random Forest used for the reported feature-importance analysis was fitted to the complete dataset, and the reported importance values were not established through an independent validation set. Consequently, the 53.3% versus 46.7% comparison should not be interpreted as evidence that TotalChanges is universally more important than PositionBias or that the two features have stable predictive importance across other samples. Independent validation is required to determine whether these relative contributions generalize.

The Isolation Forest results also require careful interpretation. Although the algorithm identified 10.3% of students as having anomalous response-position patterns and these students had lower mean scores, the analysis does not establish that the algorithm uniquely detected Positional Bias. It is possible that some of the detected patterns are associated with lower performance or other characteristics of the response data. The sensitivity analysis showed that the direction of the association remained consistent when the contamination parameter was varied from 0.05 to 0.15, but this robustness analysis does not establish the construct validity of the detected anomalies. In other words, changing the contamination parameter produced a similar direction of association, but it did not provide an external criterion against which the detected cases could be verified. Accordingly, BRAF should currently be viewed as a screening framework for unusual response-position patterns rather than a validated diagnostic measure of Positional Bias.

This study has several limitations. First, the sample consisted of 282 students from a single institutional context, which limits the generalizability of the findings. Second, the assessment included only six questions, restricting the amount of behavioral evidence available for identifying stable response-position patterns and potentially increasing the influence of individual responses. Therefore, the identified patterns should be interpreted as preliminary behavioral indicators rather than definitive evidence of persistent student characteristics.

Third, the operationalization of positional behavior has important limitations. The standard-deviation-based PositionBias indicator captures dispersion in selected option positions but cannot distinguish all possible non-random response-position structures. In addition, the Isolation Forest approach identifies unusual multivariate patterns but does not independently establish that these patterns represent Positional Bias. Thus, an external criterion or validated measure of Positional Bias was not available in the present study for direct validation of the algorithm's classifications. This limitation is particularly important when interpreting the 10.3% anomaly rate, because the contamination parameter specifies the expected proportion of observations treated as anomalies rather than providing an independently established prevalence of Positional Bias.

Fourth, the four students with more than 10 answer changes were retained in the primary analyses rather than removed as statistical outliers. A sensitivity analysis showed that the Pearson correlation between TotalChanges and final score changed only from $r = -0.1428$ ($p = 0.0164$, $N = 282$) to $r = -0.1370$ ($p = 0.0223$, $N = 278$) after excluding these four cases. The small change in the correlation coefficient ($\Delta r = 0.0057$) suggests that the observed negative association was relatively robust to these high-change observations. Nevertheless, because the number of such cases was small, their behavioral characteristics should be investigated further in larger samples.

Future research should validate BRAF using larger and more diverse samples and assessments with more items and across different academic disciplines and digital learning environments. The framework can be extended to examinations with more questions because its student-level behavioral representation is not tied to a fixed number of items. In addition, response time is already available in the present dataset and could be incorporated as a complementary behavioral feature. Other measures, such as anxiety, motivation, and additional process indicators, could also provide further insight into the mechanisms underlying the observed behaviors. Future studies should also compare the standard-deviation-based Position-Bias with alternative measures that can capture more complex response-position structures and should examine whether the same classifications are obtained across different test forms and item-ordering conditions. These directions, together with the investigation of BRAF in large-scale assessments, online learning platforms, and adaptive learning systems, may help establish its broader applicability and robustness.

Future studies should also evaluate the practical effectiveness of the proposed screening workflow rather than assuming that flagging anomalous behavior automatically leads to better educational outcomes. In particular, researchers could compare BRAF-based screening with standard instructor review and examine whether targeted follow-up interventions improve subsequent performance or reduce unproductive answer-changing behavior. The predictive validity of the framework should be evaluated using outcomes that are independent of the behavioral indicators used to construct the model. Where an appropriate binary outcome is available, measures such as sensitivity, specificity, precision, recall, and area under the ROC curve could be reported on held-out or external validation data. Such validation would help determine whether the behavioral indicators identify genuinely at-risk students or merely describe response patterns associated with performance in a particular assessment.

In addition, future experimental studies could evaluate specific interface interventions, such as explicit confirmation of revised answers, controlled opportunities for response review, or feedback on answer-changing behavior. These interventions should be tested against appropriate control conditions to determine whether they improve decision quality without discouraging beneficial answer revision. Such experiments could also examine whether the effect of an intervention differs according to students' initial performance level or behavioral profile. Importantly, these proposed interventions should be understood as testable hypotheses for future research, not as established consequences or recommendations arising from the present findings. Experimental evidence would be required to determine whether such interface modifications actually improve decision quality, reduce unproductive answer changes, or affect student performance. These studies would provide stronger evidence for translating BRAF from a behavioral screening framework into an evidence-based component of digital assessment systems.

Acknowledgment

We extend our gratitude to the students of Payame Noor University for their cooperation in this research.

References

Chen G., Liu Y. and Mao Y. (2024), Understanding the log file data from educational and psychological computer-based testing: A scoping review protocol. *PLoS ONE*, **19**(5), e0304109.

Xu X., Zhang S., Guo J. and Xin T. (2024), Biclustering of log data: Insights from a computer-based complex problem solving assessment. *Journal of Intelligence*, **12**(1), 10.

Ulitzsch E., Buchholz J., Shin H. J., Bertling J. and Lüdtke O. (2024), Using a novel multiple-source indicator to investigate the effect of scale format on careless and insufficient effort responding in a large-scale survey experiment. *Large-scale Assessments in Education*, **12**(1), 18.

Lindner M. A. and Greiff S. (2023), Process data in computer-based assessment. *European Journal of Psychological Assessment*, **39**(4), 241-251.

He S. and Cui Y. (2025), A systematic review of the use of log-based process data in computer-based assessments. *Computers & Education*, **228**, 105245.

Jiao H., Lissitz R. W. and Van Wie A. (2018), *Data Analytics and Psychometrics: Informing Assessment Practices*, Charlotte, NC, Information Age Publishing.

Embretson S. (2023), Understanding examinees' item responses through cognitive modeling of response accuracy and response times. *Large-scale Assessments in Education*, **11**(1), 9.

Ercikan K. and Pellegrino J. W. (2017), Validation of score meaning using examinee response processes for the next generation of assessments. In *Validation of Score Meaning for the Next Generation of Assessments*, New York, Routledge, pp. 1-8.

Chen L., Zhang S. and Liu J. (2026), Reducing differential item functioning via process data. *Psychometrika*, **91**(1), 95-130.

Azizi H., Amini M. S., Sulaimany S. and Mafakheri A. (2025), Visibility graph analysis for educational data: Potentials and a case study of predicting at-risk online students. *Scientific Reports*, **15**(1), 32036.

Maniyan S., Ghousi R. and Haeri A. (2024), Data mining-based decision support system for educational decision makers: Extracting rules to enhance academic efficiency. *Computers and Education: Artificial Intelligence*, **6**, 100242.

Li R. and Gao Y. (2025), Anchored answers: Unravelling positional bias in GPT-2's multiple-choice questions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1-15.

Wang S., Fellouris G. and Chang H.-H. (2019), Statistical foundations for computerized adaptive testing with response revision. *Psychometrika*, **84**(2), 375-394.

Lions S., Dartnell P., Toledo G., Godoy M. I., Córdova N., Jiménez D. and et al. (2023), Position of correct option and distractors impacts responses to multiple-choice items: Evidence from a national test. *Educational and Psychological Measurement*, **83**(5), 861-884.

Ercikan K., Guo H. and Por H. H. (2023), Uses of process data in advancing the practice and science of technology-rich assessments. In *Innovating Assessments to Measure and Support Complex Skills*, pp. 211-230.

Brady A.-M. (2005), Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, **5**(4), 238-242.

Blunch N. J. (1984), Position bias in multiple-choice questions. *Journal of Marketing Research*, **21**(2), 216-220.

Attali Y. and Bar-Hillel M. (2003), Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, **40**(2), 109-128.

Sonnleitner P., Guill K. and Hohensinn C. (2016), Effects of correct answer position on multiple-choice item difficulty in educational settings. In *Proceedings of the 10th Conference of the International Test Commission.*

Lions Maitre S., Monsalve C., Dartnell Roy P. R. G., Godoy M. I., Córdova N., Jiménez D. and et al. (2021), The position of distractors in multiple-choice test items: The strongest precede the weakest. *Manuscript submitted for publication.*

Lions S., Monsalve C., Dartnell P., Blanco M. P., Ortega G. and Lemarié J. (2022), Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review. *Applied Measurement in Education*, **35**(2), 133-152.

Fagley N. (1987), Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology*, **79**(1), 95.

Maguya A. S. (2022), Quantifying the effects of guessing, position bias and prior knowledge in multiple choice exams. *Tanzania Journal of Science*, **48**(4), 4.

Jiao C., Huang S. and Zhang Y. (2026), Embracing positional bias in multiple-choice question answering via permutation equivariant neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Bauer D., Kopp V. and Fischer M. R. (2007), Answer changing in multiple choice assessment: Change that answer when in doubt – and spread the word! *BMC Medical Education*, **7**(1), 28.

Merry J. W., Elenchin M. K. and Surma R. N. (2021), Should students change their answers on multiple choice questions? *Advances in Physiology Education*, **45**(1), 182-190.

Ferguson K. J., Kreiter C. D., Peterson M. W., Rowat J. A. and Elliott S. T. (2002), Is that your final answer? Relationship of changed answers to overall performance on a computer-based medical school course examination. *Teaching and Learning in Medicine*, **14**(1), 20-23.

Liu F. T., Ting K. M. and Zhou Z.-H. (2008), Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, IEEE, pp. 413-422.

Wilcox R. R. (2012), *Introduction to Robust Estimation and Hypothesis Testing*, Third Edition, Amsterdam, Academic Press.

Vispoel W. P., Clough S. J. and Bleiler T. (2005), A closer look at using judgments of item difficulty to change answers on computerized adaptive tests. *Journal of Educational Measurement*, **42**(4), 331-350.

Coffey J. S., Maruca A. T., Polifroni E. C. and Snyder M. (2024), Changing test answers: A scoping review. *Nurse Education Today*, **133**, 106052.

Guo H., Johnson M., Ercikan K., Saldivia L. and Worthington M. (2024), Large-scale assessments for learning: A human-centred AI approach to contextualizing test performance. *Journal of Learning Analytics*, **11**(2), 229–245.

Arslan B., Lehman B., Keehner M. and Lindner M. A. (2026), Leveraging process data to investigate the interplay between response formats and cognitive processes in digital assessment and learning environments. *Contemporary Educational Psychology*, 102461.

Noorani S. F., Mohammadi M. R., Karimi M. and Taherkhani N. (2024), CheatingRank: A Multi-Stage Approach for Detecting Cheating in Online Assessments. *Journal of Data Science and Modeling*, **3**(1), 175–199.

Cronbach L. J. and Meehl P. E. (1955), Construct validity in psychological tests. *Psychological Bulletin*, **52**(4), 281–302.

Messick S. (1995), Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**(9), 741–749.