iejee

International Electronic Journal of Elementary Education
March 2025, Volume 17, Issue 2

# Decoding Student Insights: Analyzing Response Change in NAEP Mathematics Constructed Response Items

Congning Ni[a,*], Bhashithe Abeysinghe[b], Juanita Hicks[c]

Received : 13 November 2024
Revised : 22 January 2025
Accepted : 1 March 2025
DOI : 10.26822/iejee.2025.375

[a,*] **Corresponding Author:** Congning Ni, Vanderbilt University, Nashville, TN, USA
E-mail: congning.ni@vanderbilt.edu
ORCID: https://orcid.org/0000-0001-6950-6948

[b] Bhashithe Abeysinghe, American Institutes for Research, Arlington, VA, USA
E-mail: babeysinghe@air.org
ORCID: https://orcid.org/0009-0006-4107-8615

[c] Juanita Hicks, American Institutes for Research, Arlington, VA, USA
E-mail: jhicks@air.org
ORCID: https://orcid.org/0000-0002-4906-3083

## Abstract

The National Assessment of Educational Progress (NAEP), often referred to as The Nation's Report Card, offers a window into the state of U.S. K-12 education system. Since 2017, NAEP has transitioned to digital assessments, opening new research opportunities that were previously impossible. Process data tracks students' interactions with the assessment and helps researchers explore students' decision-making processes. Response change is a behavior that can be observed and analyzed with the help of process data. Typically, response change research focuses on multiple-choice items as response changes for those items is easily evident in process data. However, response change behavior, while well known, has not been analyzed in constructed response items to our knowledge. With this study we present a framework to conduct such analyses by presenting a dimensional schema to detect what kind of response changes students conduct and how they are related to student performance by integrating an automated scoring mechanism. Results show that students make changes to grammar, structure, and the meaning of their response. Results also revealed that while most students maintained their initial score across attempts, among those whose score did change, factor changes were more likely to improve scores compared to grammar or structure changes. Implications of this study show how we can combine automated item scoring with dimensional response changes to investigate how response change patterns may impact student performance.

**Keywords:**

Response Change, Process Data, Constructed Response, Automated Scoring, Writing Behavior

## Introduction

The National Assessment of Education Progress (NAEP) serves as a critical metric providing valuable insights into student achievement across various subject areas (Johnson, 1992). With a representative sample of students nationwide, NAEP offers comprehensive statistics and reports on academic progress of the student population.

NAEP assessments cover multiple subjects and are conducted across different grade levels. In a typical NAEP

assessment, students will receive two cognitive blocks, each with a 30-minute time limit (or up to 90 minutes for students with extended-time accommodation). Students can navigate through the assessment items, within each block, in the order they are presented or via the navigation bar. Students can also revisit any item within the current block (National Center for Education Statistics, n.d.). The NAEP assessment consists of different item types (e.g. multiple-choice, drag-and-drop, constructed response) and the required mechanism(s) to answer each of these item types may be different. For example, for a multiple-choice question (Smith, 2017) a student will simply select an answer choice, but for a constructed response item (Kloosterman et al., 2015), a student must formulate and type their response. Students may also change their response to any item as many times as they like if time allows.

Student actions within the assessment are logged by the assessment system and these data are called process data (NAEP Process Data, n.d.). Behavior analysis, such as response change, can be conducted post-hoc using process data; thus, response changes for many item types such as multiple-choice and drag-and-drop, can be easily tracked since these items allow students to perform a limited set of actions. For example, in drag-and-drop items, a student is allowed only to drag components from a source to a destination. In contrast, constructed response items present a more complex scenario. A student may type their response, but by adding or deleting characters a student may conduct spelling changes, rephrase a sentence, or restructure entire sentences, which may also change the meaning of their original response. For example, a student might change "He go to school" to "He goes to school" (a grammar change) or modify "The cat sat on the mat" to "On the mat sat the cat" (a structural change). Unlike the limited actions in multiple-choice and drag-and-drop items, modifications for constructed response items are not easily visible in process data (Ivanova & Michaelides, 2023), presenting a unique challenge in exploring response change behavior for this item type.

An advantage that allows response change behavior to be observed easily in multiple-choice items and other item types is the ease of verification of the response choice. In multiple-choice items (Moore et al., 2021), given the answer key, items can be easily scored. When a student changes responses, it can be easily validated to a correct/wrong response. With this, it is also possible to investigate a student's performance gain/loss due to the response change. With constructed response items, this is not as trivial, as responses are typically graded by humans or machine scored, and changes in constructed responses are not as easily or quickly examined.

The objective of the current study is to develop a comprehensive pipeline, capable of analyzing response changes in constructed response items and categorizing them into dimensions to gain a better understanding of their impact on student performance, student behavior, and learning mechanisms.

## Literature Review

In this section, we will explore prior research that has at least tangential relationships with the investigation we are conducting into student editing and response change behaviors in constructed response items. First, we look at the current state of general response change literature as this is the first work investigating such student behavior. Then, we draw inspiration from student writing and editing research to prepare the background of our current investigation into response change for constructed response items.

### *General Response Change*

Response change or answer change behavior refers to the modifications that students make to their answers during an assessment (van der Linden & Jeon, 2012; Tiemann, 2015). Understanding these changes is crucial, as it provides insights into cognition and assessment strategies. Prior work has explored student response change behavior in standardized paper-pencil assessments. However, with the advent of digital assessments, process data has become a valuable resource for analyzing response change behavior. Process data includes timestamps and interaction logs that provide detailed records of student behavior during an assessment. This data allows researchers to study not just the final answer but also the sequence of actions leading to it (Ercikan et al., 2020).

In process data, intra-visit changes involve changing an answer before moving on to another question, while inter-visit changes occur when students revisit a question to revise their answers. As defined by Ouyang et al. (2019a), changes within the same visit could be due to typographical errors or immediate corrections and are generally not considered response changes. In this study, we focus on inter-visit changes. These changes provide insight into how students rethink and re-evaluate their previously written responses. This distinction allows us to understand the cognitive processes involved in checking and modifying responses better. Previous studies have demonstrated the significance of studying inter-visit changes to gain insights into student learning and behavior (Qiao & Hicks, 2020; van der Linden & Jeon, 2012).

Since inter-visit changes reflect a deeper engagement with the problem-solving process, prior research has primarily examined these behaviors in multiple-choice questions (MCQs). The structured nature of MCQs allows researchers to track response

changes efficiently, as process data capture distinct answer selections, and verification of correctness is straightforward (Qiao & Hicks, 2020). Consequently, research into response change patterns such as right to wrong (RW), wrong to right (WR), right to right (RR), and wrong to wrong (WW) is common (van der Linden & Jeon, 2012). These patterns help understand the impact of answer changes on performance. For example, McMorris et al. (1991) found that high-ability students were less likely to change their initial answers; but when they did, their answer changes were mostly from incorrect to correct. Research also shows that students often benefit from changing their responses which improve their score (Bridgeman, 2012; Tiemann, 2015). Liu et al., (2015) used GRE data to explore response change patterns and found that students with higher abilities benefited more from response changes. Similarly, studies have noted the effect of item difficulty on response change behavior, with easier items having more frequent WR changes and harder items showing more WW changes (Al-Hamly & Coombe, 2005; Jeon et al., 2017; van der Linden & Jeon, 2012; Tiemann, 2015).

*Response Change in Constructed Response Items*

In constructed response items (CR), students write their own responses instead of selecting from a given set of options. This presents two unique challenges in observing response changes. First, in process data, response modifications to constructed response items are recorded at the character level, meaning that each insertion or deletion of a character is logged individually. However, in reality, students often revise entire words or phrases, which can change the overall meaning of their response. Second, there is no direct mechanism to validate students' intermediate responses (i.e., responses which come before the final response – NAEP response data includes correct or incorrect scores for a given item, but this is only for the final response). This complexity requires a nuanced approach to categorize and understand these changes. Unfortunately, the literature on response change for constructed response items is scarce as this has not been previously analyzed with respect to constructed response items (Benjamin et al., 1984; Jeon et al., 2017; Qiao & Hicks, 2020; van der Linden & Jeon, 2012); therefore, we draw inspiration from writing and editing literature to help support the foundation for the current research.

Research in assessment writing and CR items has demonstrated that students frequently make changes during the assessment process. These changes can significantly impact on the quality and correctness of their responses. For example, Engblom et al. (2020) found that students often revise their responses, particularly focusing on spelling corrections prompted by software indicators. This indicates active engagement in improving their responses through various modifications such as grammar corrections and sentence restructuring. Tate & Warschauer (2019) examined digital writing assessments and found that keypresses and mouse clicks provided valuable data on student writing processes, revealing patterns that correlated with writing performance. They also highlighted that digital writing involves different cognitive processes compared to traditional writing, including frequent revisions and modifications (Hojeij & Hurley, 2017).

Kim & Kim (2022) investigated student responses in large-scale assessments, categorizing answers into correct, partially correct, and various error types. They found that higher-achieving students tend to make fewer errors compared to lower-achieving students. A similar observation was also made by Liu et al. (2015). Despite the limited direct research on response changes in CR items within assessments, the studies from writing research may offer a framework to understand and analyze the modifications students make in constructed response items. To reiterate, these are the core concepts that we draw from the writing and editing literature:

1. when constructing their responses students may make revisions, focusing on particular modifications (Engblom et al., 2020),

2. revisions can be observed by keystrokes and mouse clicks, providing insights into various patterns related to writing performance (Tate & Warschauer, 2019),

3. students will self-edit hoping to improve their own writing (Hojeij & Hurley, 2017).

*Purpose of Current Study*

Students' writing patterns in CR items, such as adding or removing words, correcting spelling errors, and restructuring sentences are not easily captured. Therefore, analyzing response changes in CR items presents many challenges from data capture to analysis compared to other item types that have been previously researched.

Following the prior work on writing and editing, we aim to explore students' response changes in CR items by categorizing various text changes into dimensions (dimensional changes) such as grammar, structure, and factor. Grammar changes involve spelling or grammatical corrections, structure changes involve reordering or modifying sentence structures, and factor changes involve changing the conceptual meaning of the response. We then use a classification model to investigate the effects of these dimensional changes on student scores.

By analyzing how students change their responses in CR items we hope to reach two goals: 1) address the

research gap of CR items response changes as well as the gap of a general analysis of CR items, and 2) propose a framework which can be used to analyze CR items in terms of student writing and editing. Through this process, we hope to analyze specific changes in CR items which extend further than the typical research into character addition/deletion. By exploring these dimensions, we aim to provide deeper insights into how students' response change behavior in CR items might be related to their testing behavior, performance, and learning processes.

**Research questions**

In our study, we aim to understand and analyze the dimensional changes in students' constructed responses. Our framework is designed to address two primary research questions and outline future work:

RQ1: How can we categorize response changes in students' constructed responses across multiple visits?

RQ2: Can we develop an item scoring model to score each visit response and analyze the relationship between dimensional changes and score changes?

**Methodology**

*Data*

Data for this study come from the 2022 NAEP Grade 8 mathematics assessment. Specifically, we targeted item 7 from block MB which contains 15 items of different types (e.g., Multiple-Choice, Extended Constructed Response, Drag and Drop, etc.). Item 7 is a short-constructed response (SCR) item focusing on algebra. It is a multi-part, hard-difficulty item that poses a question about the intersection of two distinct lines in an xy-plane. Students are tasked with responding to a multiple-choice question and explaining their reasoning in a short-constructed response format (Figure 1). Item 7 provides a concise yet structured format for analyzing response changes and this item type allows us to systematically categorize different types of modifications (e.g., grammar, structure, factor) while ensuring a manageable scope for analysis. A total of 13,300 students were used in this analysis. A small group of students conducted revisits and further generated response changes. This group contains approximately 400 students (3%) from the block.

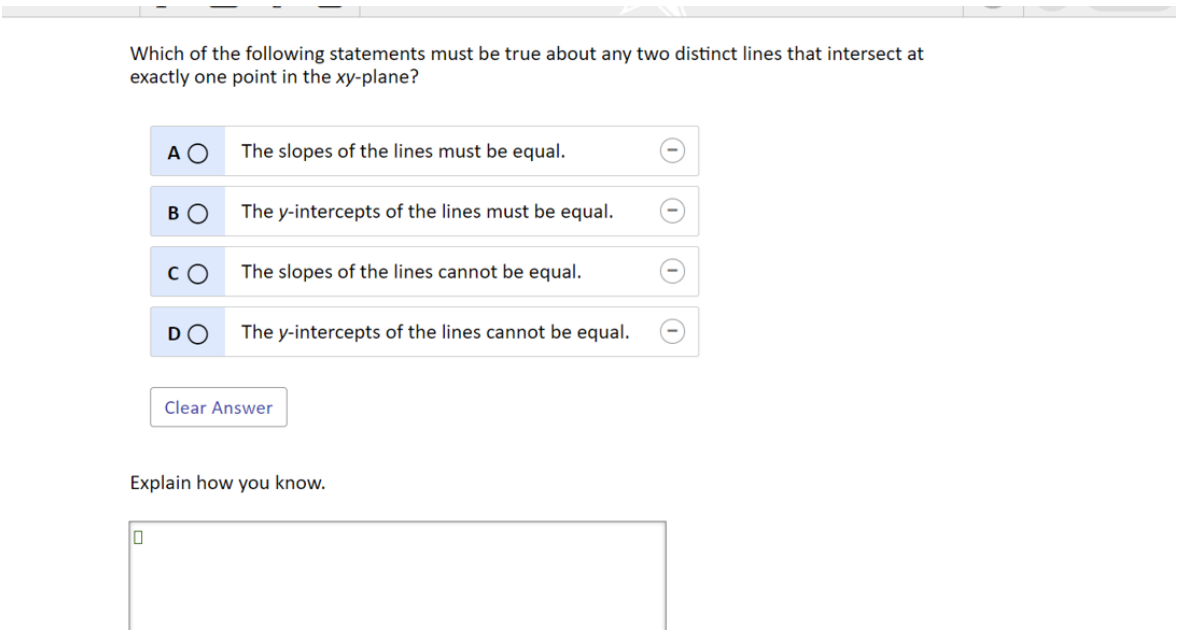
Sample Correct Responses

- Correct Selection: C. The slopes of the lines cannot be equal.
- Explanation: The slopes cannot be equal because if they were equal, the lines would be parallel. Distinct parallel lines do not intersect.

Scoring

- Correct: Correct selection with an acceptable explanation.
- Partial: Correct selection with a partially acceptable explanation or an incorrect selection with an explanation that supports the correct selection.
- Incorrect: Correct selection with an unacceptable or no explanation; or an incorrect response.

**Figure 1.**
*Item 7 screen capture from eNAEP.*




SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

*Data Processing*

Responses to constructed response components are captured for each keystroke as an event and responses to multiple-choice items are captured as a numerical entry representing the option choice (i.e., 1-A; 2-B; 3-C; 4-D) in process data. The accumulation of individual keystroke events creates the full response as typed by the student. Therefore, process data is rich in information on which we can conduct various analyses. For item 7, the student's final response contains both the multiple-choice response and the constructed response. Using a combination of text processing techniques, each response can be converted into plain-text format. The result of data processing for item 7 is an extended dataset that includes cleaned (e.g., deduplicated data) and organized (e.g., data ordered by timestamps) student responses, incorporating both the multiple-choice response and extracted plain text for each item visit. The data is grouped by student to maintain the sequence of response changes made by each student, ensuring a comprehensive view of their behavior throughout the assessment process. This is the dataset that will be used for analysis of both RQ1 and RQ2.

*Analysis Plan*

The goal of this research is to explore and operationalize the response change concept for constructed response items. To do this, we have introduced procedures on what establishes a response change for a CR item and then further categorize the response changes into dimensions. The dimensional analysis of response change offers several benefits for educational assessment. It provides a structured mechanism to capture and analyze the complexity of student responses, allowing for a more nuanced understanding of their behavior. Moreover, dimensional analysis can enhance the reliability and validity of assessment scores and interpretations of scores by accounting for the various types of changes students make. This method can also help detect potential issues such as misunderstanding of the task, misconceptions, or lack of knowledge, providing valuable feedback for both students, educators, and researchers.

*Definitions*

To help operationalize response changes in constructed response items we provide definitions for aspects of student behavior that support response change.

- Visit: Each entry into an item, performing any action, and then exiting the item.
- Response Change: When a student revisits an item and modifies their previous response. This can occur multiple times and includes any alteration made to the initial response.
- Dimensional Change: A specific type of modification within a response change, referring to a meaningful alteration(s) that affects different aspects of the response.

Additionally, we provide examples of each type of response change found in student responses to item 7. The response changes are then aligned with the dimension that best describes the response change (Table 1).

*Introduction of Study Framework*

This study introduces a framework (Figure 2) that ties together the two research questions and allows us to examine response changes in constructed response items, explore how these changes are related to dimensions of change, and investigate

**Table 1.**
*Dimensions of Response Change.*

| Response Change Type | Example | Dimension |
|---|---|---|
| Misspellings | Correcting "recieve" to "receive". | Grammar Change |
| Punctuation | Adding a period at the end of a sentence. | |
| Capitalization | Changing "john" to "John". | |
| Verb Tense | Changing "He go to school" to "He goes to school". | |
| Stemming | Changing "running" to "run". | |
| Word Choice | Replacing "happy" with "joyful". | Structure Change |
| Concision | Changing "In my opinion, I think that" to "I think that". | |
| Sentence Reordering | Changing "He ran quickly to the store" to "Quickly, he ran to the store". | |
| Paragraph Reorganization | Changing the order of sentences or paragraphs for better flow. | |
| Changes in Meaning | Changing "He goes to school" to "He headed home". | Factor Change |
| Elaboration | Expanding "The cat sat on the mat" to "The small, fluffy cat sat comfortably on the mat". | |
| Detail Removal | Removing redundant or irrelevant information to streamline the response. | |

**Figure 2.**
*Analysis plan and framework proposed in this study.*

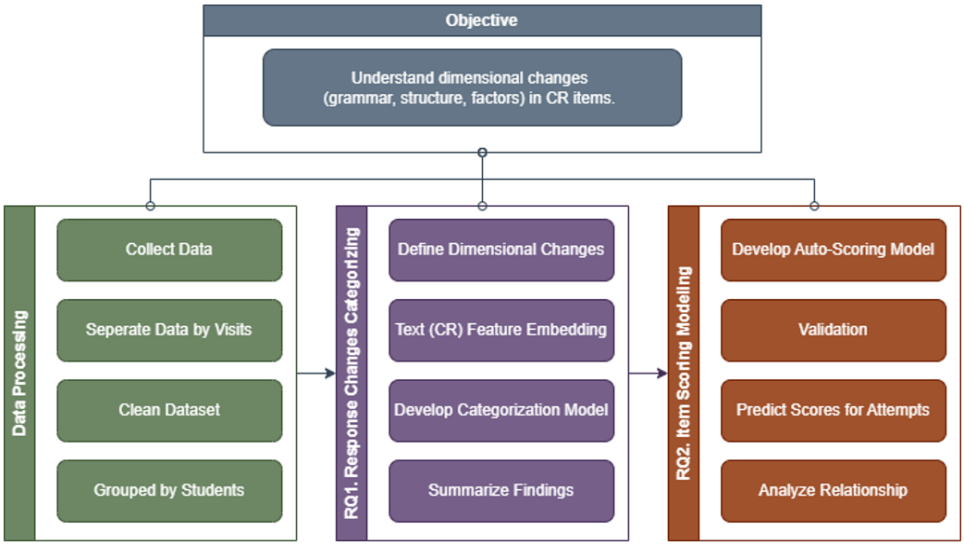
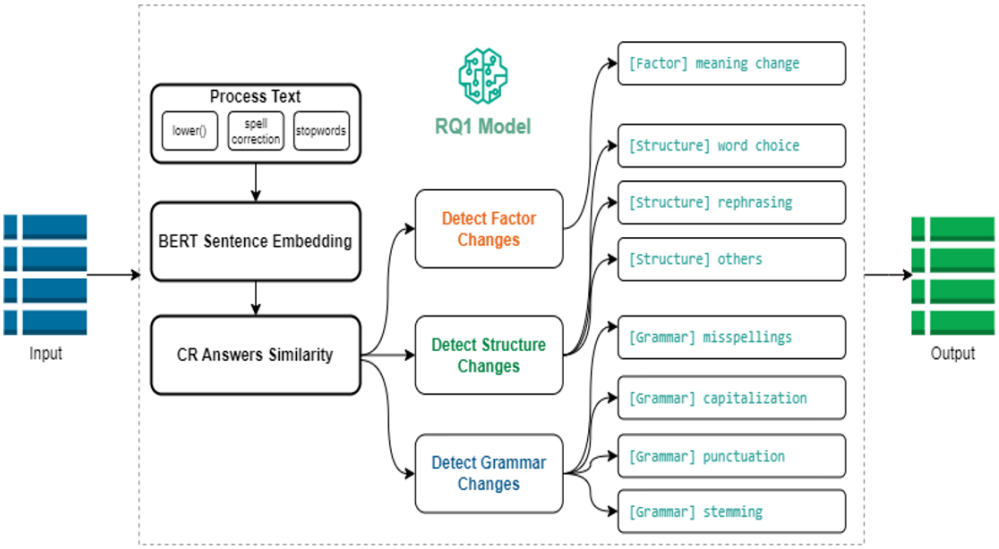



**Figure 3.**
*Model of the process developed for RQ1.*




how these dimensions of change are related to student scores. The data processing stage highlights the steps necessary to prepare the data for analysis in RQ1 and RQ2. The stages for RQ1 and RQ2 highlight the process of responding to each research question by categorizing student response changes and scoring responses, respectively. Improvements to the framework are anticipated, which is the reason for modular implementation. We plan to refine our models and methodologies based on the findings from RQ1 and RQ2. The versatility of this framework lies in its ability to be adopted to analyze similar behavior in other constructed response and text-based items.

RQ1: Dimensional Changes Categorization

A simple illustration of the RQ1 model process is available in (Figure 3). The input for the model consists of pairs of constructed responses (pre-response change and post-response change) from students. These pairs of responses are processed to detect the changes made between attempts. Responses are converted into sentence embeddings using BERT, which captures the semantic meaning of the responses (Devlin et al., 2019). The processed responses are then compared for similarity to detect changes.

To measure the similarity between sentences, we compute the cosine similarity between the embeddings of the pre-response change and post-response change. Cosine similarity is a metric that quantifies the degree of similarity between two vectors by measuring the cosine of the angle between them. A value close to 1 indicates high similarity, meaning the response remains largely unchanged in meaning, whereas a value closer to -1 suggests a

significant difference in content. This similarity score helps to identify changes that are not immediately obvious from a simple text comparison. High similarity indicates that responses are semantically similar, whereas low similarity suggests significant changes. To effectively categorize the response changes, we adopt a hierarchical structure. Factor changes take precedence, followed by structure changes, and then grammar changes. This approach ensures that significant changes in meaning are identified first, followed by changes in structure, and finally minor grammar changes. These changes are determined based on predefined linguistic rules and manual reviews as described later in this section.

*Dimensional Change Detection*

Factor changes represent significant changes in the underlying meaning of the text. The input for detecting factor changes is the fully preprocessed text, including lemmatization and removal of stop-words. This ensures that the analysis focuses on the core content and meaning of the responses. The model detects factor changes by measuring the overall semantic similarity between the pre-response change and post-response change responses. Low similarity, in our case less than 0.85, indicates a factor change, suggesting a shift in the conceptual understanding or approach to the problem. This threshold was determined through an empirical review of manually annotated response changes, where we analyzed the distribution of similarity scores and identified 0.85 as a point that effectively distinguished meaning-altering modifications from minor edits. The process involves tokenizing the text and extracting unique words, which are then compared using BERT embeddings. The output includes notes on the specific factor changes detected, such as "[Factor] meaning change."

Structure changes involve modifications to the arrangement of words and sentences while preserving the original meaning. The input for detecting structural changes is the preprocessed text, where words are lemmatized, but stop-words are retained. This helps to focus on the core structure of the sentences. The model detects structural changes by comparing the semantic similarity of sentence embeddings. High similarity with a different word order or rephrasing indicates a structural change. In our case, similarity scores greater than 0.95 indicate a structural change. This threshold was determined by manually reviewing 50 samples. The detection process includes splitting the text into sentences and identifying common and unique sentences between the pre-response change and post-response change responses. The unique sentences are then compared using text embeddings to measure their similarity and changes in word choice and sentence reordering are identified. The output includes detailed notes such as "[Structure] word choice" or "[Structure] rephrasing."

Grammar changes focus on spelling, punctuation, capitalization, and stemming. The input for detecting grammar changes is the original text without any preprocessing for spelling correction or stop-word removal. This allows us to identify raw grammar errors and changes. The process of detecting grammar changes involves several steps. First, the text is tokenized – this is the process of breaking the text down into smaller units (tokens); in our case, we use the word tokenize[1], and each token is checked for spelling errors. Differences in punctuation are identified by analyzing the counts and positions of punctuation marks. Capitalization changes are detected by comparing the case of words between pre-response change and post-response change responses. Stemming changes are identified by comparing the lemmatized forms of words to detect changes in word forms. Finally, the output includes detailed notes on the specific grammar changes detected, such as "[Grammar] misspellings" or "[Grammar] punctuation." The categorization of these dimensions may offer insights into how students modify their responses across multiple visits (Ouyang et al., 2019).

*RQ2: Item Scoring*

The item scoring model used to address RQ2 is shown in Figure 4. This model employs a multi-step classification approach to evaluate student responses during the revision process. We use logistic regression for both primary and secondary classifications, as we observed varying model performance when using other classification methods. Early experiments showed good performance with logistic regression. This method aims to accurately predict the scores for each visit response based on both the multiple-choice response and constructed response.

*Item Scoring Training & Fitting*

The input for this model is the resultant dataset of the data preprocessing section, which includes both the multiple-choice response and constructed response for each student. The model does not include information on the input being an intermediate or final response. The constructed response is preprocessed using text normalization techniques, including lowercasing, removal of stop-words, and text vectorization using TF-IDF (Aninditya et al., 2019). The response choice is one-hot encoded to create a numerical format suitable for machine learning models. The feature embedding, which is the input of the model, consists of the preprocessed constructed responses and the one-hot encoded response choice. The feature embedding is then put into a matrix. The combined matrix is then used for both primary and secondary classifications. The primary classification predicts whether a response is "Incorrect" or "Not Incorrect". For responses classified as "Not Incorrect," a secondary logistic regression model further classifies

**Figure 4.**
*Model of the process developed for RQ2.*

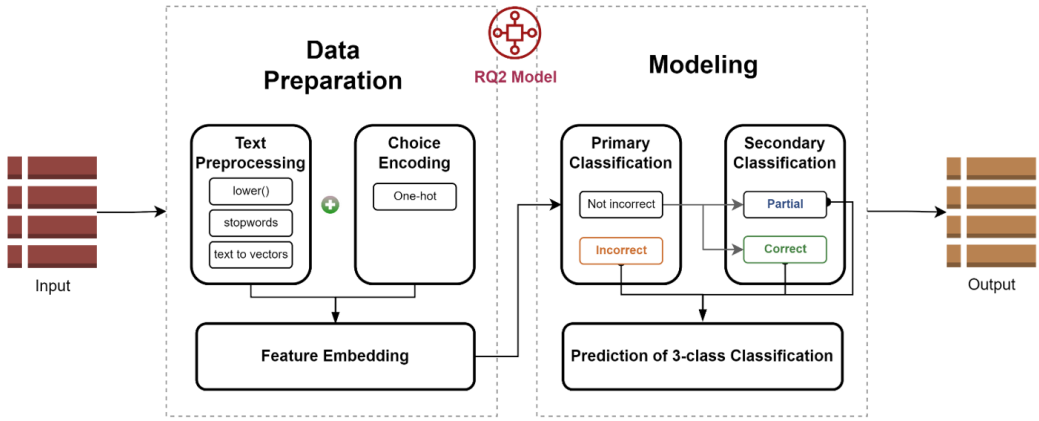
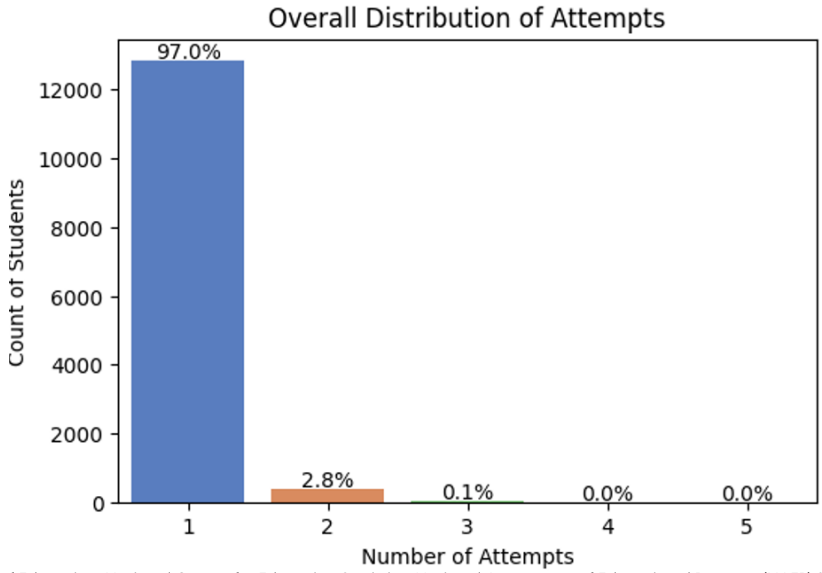



**Figure 5.**
*Number of students and their attempts to the selected item.*




SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

them into "Partial" or "Correct." The ultimate output of the model is a predicted score for each response attempt, indicating whether the response is "Incorrect," "Partial," or "Correct."

During training, logistic regression models are fitted with a maximum of 1000 iterations. The cross-validation process involves splitting the data into five stratified folds, maintaining the same ratio of each class in each fold. This ensures that each fold has a proportional representation of the different classes. The primary classifier is trained on the binary classification task (0 for "Incorrect" and 1 for "Not Incorrect"), and the secondary classifier is trained on the binary classification task (0 for "Partial" and 1 for "Correct") for responses predicted as "Not Incorrect." To handle the class imbalance in the secondary classification, we resampled the minority class ("Partial") to match the size of the majority class ("Correct").

After evaluating the model performance using cross-validation, the model is trained on the entire training dataset to generate the final model for future predictions. This ensures that the model is trained on all available data to maximize its predictive accuracy. The trained models, along with the vectorizer and encoder, are saved for future use, enabling the application of the model to new data. The final model is then applied to intermediate attempts to obtain "temporal scores," reflecting student performance at different stages of their response modification process.

By analyzing the relationship between the predicted scores and the dimensional changes detected in RQ1, we aim to gain a deeper understanding of how changes in student responses impact overall student performance.

## Results

This work analyzes 13,300 students who participated in block MB of the 2022 NAEP mathematics assessment. All students who are included in the sample attempted the selected CR item. Results reveal that

many students do not revisit an item once they have completed their initial response. However, we did find a small group of students who conducted revisits and response changes. This analytical sample resulted in approximately 400 students (~3%). Results show that the average number of item attempts per student is 1.06, indicating that repeated attempts are relatively uncommon among students. The maximum number of attempts recorded is 5 (Figure 5), highlighting a small group of students who exhibit more persistent engagement. Focusing on the behavior of students who make multiple attempts, we aim to uncover strategies related to response changes that can be used to support students in improving their problem-solving skills and learning outcomes.

*RQ1: Dimensional Changes in Student Responses*

In the methodology section we introduced a process to categorize response changes into dimensional changes for constructed responses. The model essentially categorizes response changes into three dimensions: grammar, structure, and factor. The application of this process revealed that each attempt to answer could involve multiple dimensional changes. Specifically, the number of dimensional change types per attempt were distributed as follows: approximately 260 attempts involved two types of change, over 70 involved one type of change, and over 70 attempts involved three types of changes.

**Dimensional Changes**

The grammar change dimension includes misspellings, punctuation errors, capitalization inconsistencies, verb tense changes, and stemming differences. Analysis showed that misspellings were corrected by students in approximately 200 instances. Punctuation changes were observed in 180 instances, capitalization changes observed in 150 instances, and stemming changes were observed the least, in about 20 occurrences (Figure 6a).

The structure change dimension describes modifications to the arrangement of words and sentences while preserving the original meaning. Many structure changes fell into a broad "other" category and about 10 instances involved sentence rephrasing. Results also showed that changes in lexical choices were not conducted significantly (Figure 6b).

The factor change dimension refers to a significant shift in the underlying meaning of the response. Results identified about 360 instances of meaning change, highlighting a substantial area where students altered their conceptual understanding or approach to the item. Since the factor change dimension does not have subcategories requiring breakdowns like grammar and structure, a separate figure was unnecessary, as it would contain only a single bar.

**Figure 6a.**
*Number of students with various types of grammar changes.*

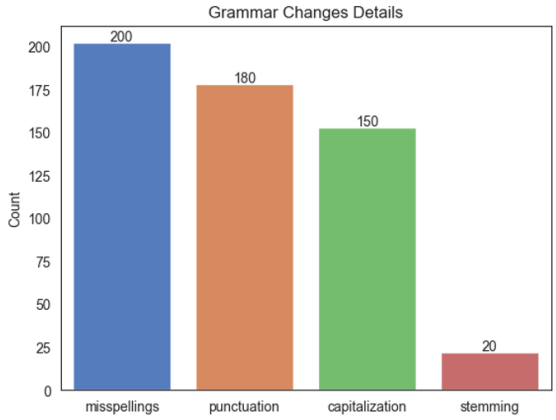
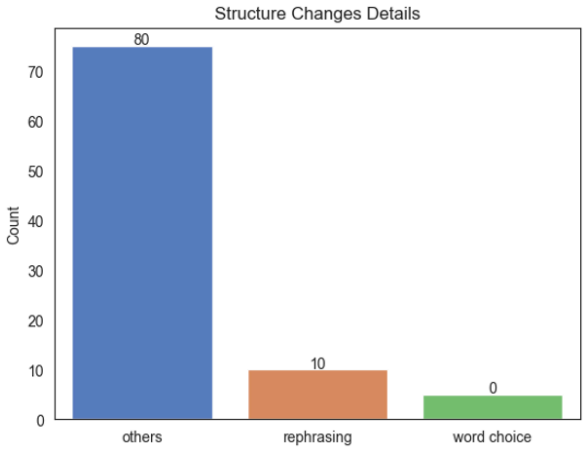



**Figure 6b.**
*Number of students with various types of structure changes.*




SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

*Demographic Analysis of Dimensional Changes*

The dimensional changes were further analyzed across various demographic categories to understand the patterns and disparities among different student groups. Moreover, Fisher's exact tests were conducted to examine if the differences between groups were statistically significant. Figure 7 reports the ratios of students who conducted a dimensional change given that a response change was conducted. The dimensional changes were normalized by using the ratios to ensure an accurate representation of each group.

Structure changes were slightly more prevalent among female students (23.7%) compared to male students (19.3%) but we found that the difference was not statistically significant (Figure 7a). Racial groups showed varying patterns in dimensional changes (Figure 7b). Factor changes were the most common among all racial groups, even though they were the least common among White students (83.2%).

Results from the Fisher's exact test showed that these differences were not statistically significant. However, there were significant differences in the race category for structure changes, such that students from all races were less likely to conduct structure changes (p = .01).

Students with an Individualized Education Program (IEP) and identified as having a disability (SD) showed slightly higher percentages of grammar (86.7%) and factor changes (90%) compared to students without IEPs and identified as not having a disability (Figure 7c); these differences were also noted as not significant. Similarly, English Learners (EL) had a slightly higher percentage of grammar changes (88.9%) and factor changes (87.6%) compared to non-El students (Figure 7d) which was also found to be not significant. However, Fisher's exact test (p = .01) indicated that non-EL students were more likely to conduct structure changes (23.2%) compared to EL students (5.6%).

Students who were eligible/ineligible for Free/Reduced-price lunch eligibility (Figure 7e) also showed varying ratios for dimensional changes that were not statistically significant. Overall, the most variation among demographic groups was for structure changes. The detailed breakdown of dimensional changes and their distribution across demographic groups provide a comprehensive understanding of student behavior and learning process in the assessment context.

*RQ2: Item Scoring Model*

For RQ2, we implemented a dual-layer classification model using logistic regression for both primary and secondary classifications. This model was trained to predict student scores based on student written responses and student multiple-choice selection data. The performance of the model was evaluated using accuracy and classification reports. The results are summarized in Table 2.

**Table 2.**

*Model performance for predicting the score of a student response.*

| Metric | Incorrect | Partial | Correct | Macro average |
|---|---|---|---|---|
| Precision | 0.95 | 0.19 | 0.81 | 0.65 |
| Recall | 0.96 | 0.02 | 0.85 | 0.61 |
| F1-Score | 0.96 | 0.04 | 0.83 | 0.61 |
| Overall Accuracy | - | - | - | 0.92 |

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

As Table 2 shows, overall accuracy of the model is 92%. However, it is evident that the model performs exceptionally well in predicting "Incorrect" and "Correct" responses but struggles with "Partial" responses. This is likely due to the class imbalance, which was somewhat reduced by resampling the minority class in the secondary classification layer. This issue is illustrated in Table 3, where partial classifications are attributed to both incorrect and correct classes.

**Figure 7a-7e.**

*Analysis of dimensional changes by gender (7a), race (7b), individualized education program (IEP) (7c), limited English proficiency (LEP) (7d), and school lunch (7e).*

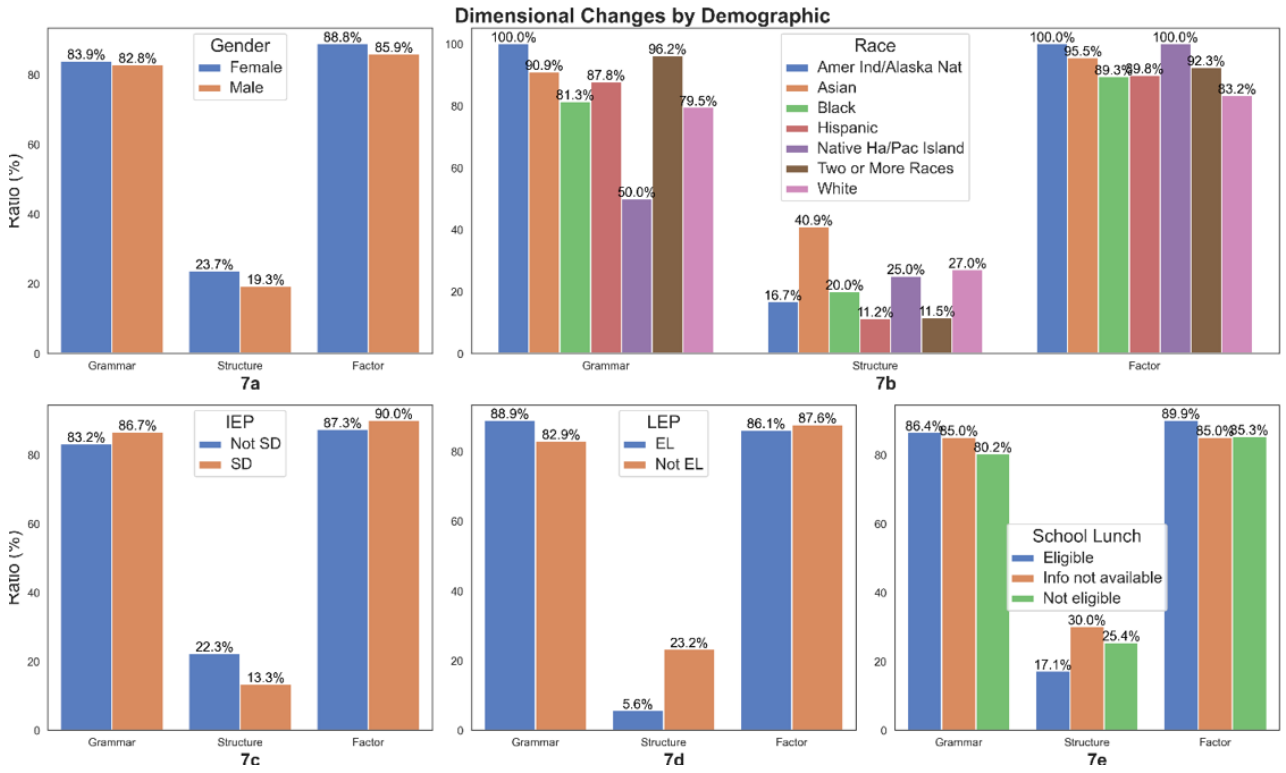


**Table 3.**
*Confusion matrix for the item scoring models performance.*

| | | Predicted | | |
|---|---|---|---|---|
| | | Incorrect | Partial | Correct |
| True | Incorrect | 10050 | 30 | 350 |
| | Partial | 170 | 10 | 150 |
| | Correct | 370 | 1 | 2070 |

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

### *Application of the Trained Model*

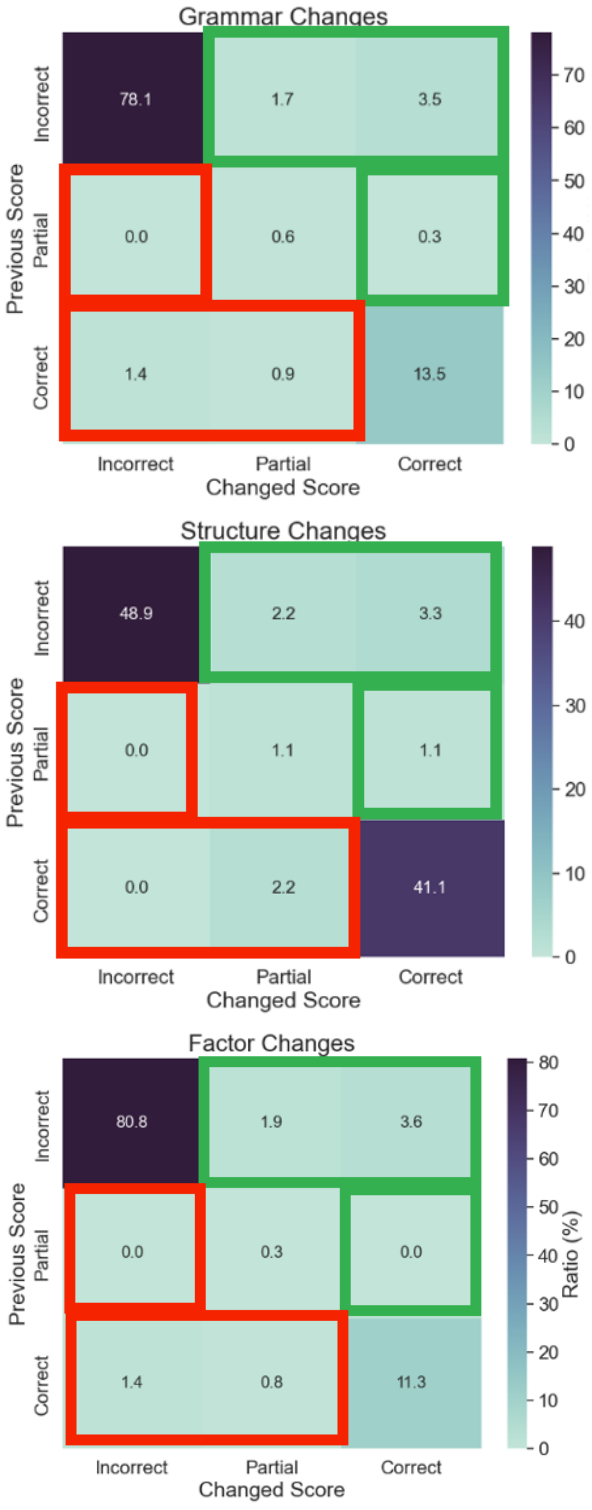
After training and evaluating, the item scoring model was applied to all attempts to generate predicted scores. These results provided insights into how students' scores changed between attempts. Given the concise nature of the written responses, it was anticipated that changes in the meaning of responses (factor changes) were more likely to result in score modifications compared to grammar or structure changes. However, results revealed that most students maintained their initial score across attempts. Among those students whose score did change, however, factor changes were more likely to improve scores compared to grammar or structure changes. The heatmaps in Figure 8a-8c illustrate the percentage of score transitions for grammar, structure, and factor dimensional changes.

Improvements and decreases in scores across all three dimensions are quite similar. We observed the best improvement in score for students conducting structure changes at 6.6% (sum of all the green boxes in Figure 8b). Grammar and factor changes improved 5.5% of student responses (the sum of all the green boxes in 8a and 8c, respectively). Structure or factor changes contributed to student score decreases 2.2% of time (sum of all the red boxes in 8b and 8c, respectively), while grammar change decreased student scores 2.3% of the time (sum of all the red boxes in 8a). Overall, more students increased their score rather than decreasing it when performing any dimensional change.

### *Demographic Analysis of the Item Scoring Model*

Because changes to the factor dimension create the most change in scores, demographic analysis regarding student performance is only shown with respect to factor changes. An examination of gender-based differences indicates that both male and female students show a moderate proportion of score improvements from "Incorrect" to "Correct" (0 to 2) following factor changes (Figure 9). Specifically, 3% of male students and 4% of female students exhibit this transition. Conversely, the shift from "Incorrect" to "Partial" (0 to 1) is less prevalent, occurring in 2.5% of female students and 1.2% of male students. As previously mentioned, a large majority of students maintained their scores across change attempts (0 to 0; 2 to 2). Overall, these observations suggest a slightly higher likelihood of score improvement among female students after making factor changes.

**Figure 8a-8c.**
*Performance of the item scoring model for grammar changes (8a), structure changes (8b), and factor changes (8c) represented in a confusion matrix heatmap.*




SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

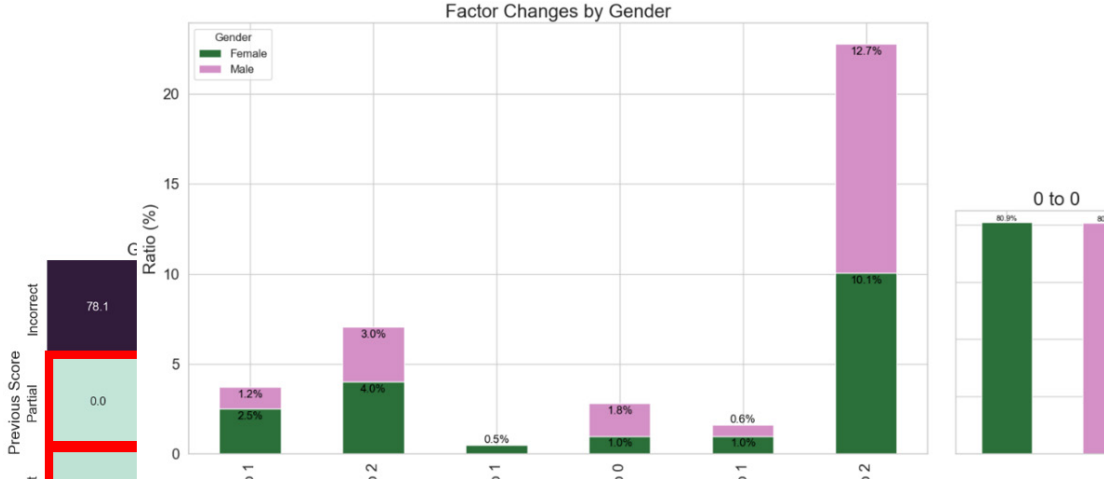
Analyzing factor changes by race reveals distinct patterns of score transitions among different demographic groups (Figure 10). Native Hawaiian/ Pacific Islander students exhibit the highest rate of improvement from "Incorrect" to "Partial" (0 to 1) at 25.0%. For the transition from "Incorrect" to "Correct" (0 to 2), students identified as Two or M ore Races display the highest rate at 8.3%, while all other groups have similar rates of transition. Among the groups maintaining their correct scores (2 to 2), Asian students stand out with the highest rate of 23.8%, followed by White students at 16.2%. Another interesting observation is that American Indian/Alaska Native students appear only to have retained their score, without improving (0 to 0). Other demographic variables (i.e., IEP, LEP, School Lunch) did not show meaningful patterns in factor changes; thus, we did not report them in this study.

## Discussion & Conclusion

In this study we analyzed the data from over 13,300 students who participated in the 2022 NAEP Grade 8 mathematics assessment. The selected item for analysis requires students to select an answer choice and then explain their reasoning. Prior research suggests that students engage in response change behaviors (e.g., Engblom et al., 2020; Jeon, De Boeck, et al., 2017; McMorris et al., 1991), some of which are positively related to problem-solving behaviors that help improve student performance (Al-Hamly & Coombe, 2005; Beck, 1978; Liu et al., 2015). Although there has been research conducted in response change behavior, to our knowledge response change analysis for constructed response items has not been conducted. Thus, our work contributes to this research area by introducing dimensional categories to analyze how students change their responses and by introducing how we can combine automated

**Figure 9.**

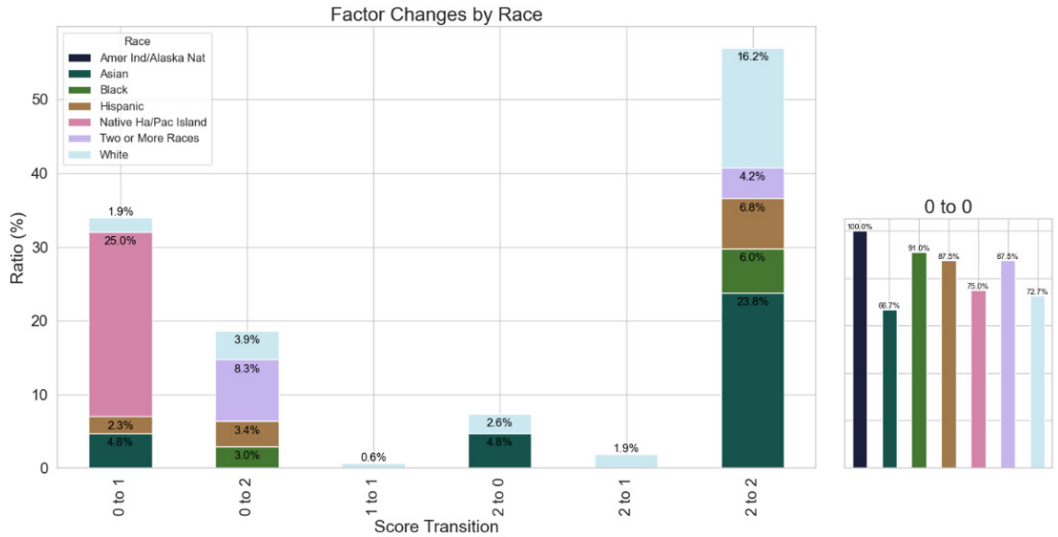
*Score transition patterns in factor dimension by gender.*




SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

**Figure 10.**

*Score transition patterns in factor dimension by race.*




SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2022 Mathematics, Grade 8.

item scoring with dimensional changes, to investigate how response change patterns may impact student performance.

To realize the above goal, two models (one algorithmic model and one machine learning model) were created to extract dimensional changes from constructed responses and then score the intermediate responses. Both models were analyzed for their accuracy and performance; thus, both models demonstrated the ability to accurately categorize dimensional changes and predict scores. Together, the components of this work encapsulated a framework to analyze constructed response items. The framework was created so that the components are loosely coupled, meaning that each component can be changed without making heavy changes to the framework itself. This makes the framework accessible for discovery of any new dimensions, while also helping to improve the item scoring model without changing the base of the framework. This framework supports the research goal by creating an end-to-end system, therefore reducing engineering challenges potentially faced by others who are interested in this framework for their research in the future.

As noted in the results section, we observed that only a small number of students conducted response changes. However, the patterns and changes within this small group of students can still provide valuable insights about student assessment behaviors. The small group size of response changing students indicates that persistent engagement in students is uncommon behavior, or at least for constructed response items. However, it is still interesting to note that students who showed engagement and changed their response were more likely to improve their score. This was highlighted in the literature (Jeon et al., 2017; van der Linden & Jeon, 2012) and also in the results we presented in the previous section. While this is an observed phenomenon in literature it seems as if the student population is yet to understand the impact of response changes could bring.

*RQ1*

Analyzing dimensional changes with respect to response change is a novel application. However, similar analyses have been conducted in other areas such as writing and editing research (Engblom et al., 2020; Malekian et al., 2019; Tate & Warschauer, 2019). Since this work analyses a constructed response item, we found that tangential research in writing and editing was helpful and helped to provide context for the results of this study. We learned from this literature that students tend to make edits/changes to their responses focusing on specific modifications, hoping that these modifications would improve their score (Engblom et al., 2020; Hojeij & Hurley, 2017). We formulated the dimensional categorization on this premise and analyzed how students make changes.

The findings of the dimensional categorization process are interesting. Overall, we observed that students tend to conduct more grammar changes which is parallel to the findings of Engblom et al. (2020). Simple changes such as spelling fixes and punctuation are visible and easy to conduct. Structure and factor changes require increased effort from the student and since the item is at the second half of the assessment this may be a reason for that behavior to be displayed less (Lee & Jia, 2014; Pools & Monseur, 2021; Setzer et al., 2013). However, when observing the demographic breakdown of the dimension results, results show that grammar and factor changes are the most used categories of response change. Another interesting observation from the results of RQ1 was the disparity in structure changes between English Learners (EL) and non-EL. Non-EL students showed evidence of structure change nearly four times more than EL students. Modifications to the arrangement of words and sentences made by non-EL students, while preserving the original meaning of their responses, potentially suggests that non-EL students have a stronger command of the language.

*RQ2*

In general response change literature, for other item types, researchers tend to explore how the change itself will impact the students score (scoring only the final attempt). However, with NAEP process data we can extract the intermediate responses using process data as well as the final scored response. Obtaining the score for intermediate responses is not trivial. For a multiple-choice item it is a matter of validating the intermediate choice against the answer key. However, for constructed response items, validating the intermediate response is not a straightforward process. Therefore, to obtain intermediate scored responses we trained a machine learning model.

The scoring model is a logistic regression model trained with a few natural language processing features which we engineered for this study. While evaluating the model, we noted that the model performed greatly in predicting incorrect and correct scores, but with partial scores, the model struggled possibly due to class imbalance. The issue of class imbalance is a difficult issue which in some cases can be solved via resampling or data augmentations (Chawla et al., 2002). We oversampled for the minority class; however, we did not see improvements in our model for the partial score group. The most likely reason in such cases is that either the model is too simple or the features that are fed into the model are not comprehensive enough to capture the underlying patterns. While it is true that this is a simple model, the features also could have contributed to the decreased performance in the partial class.

Even with such challenges we were able to still employ the model to extract student scores on

dimension and response changes. While there were multiple dimensional changes observed that impacted student performance, factor changes - which involve changes in the meaning of responses - were particularly influential in leading to changes in scores. As mentioned before, grammar and structure changes do constitute as changes; however, they may not change the response in terms of conceptual understanding. Changes to the factor dimension are highly impactful since they effectively change the meaning of the response. To conduct this change, a student may need to change their comprehension of the question or recall new ideas and facts that would change their understanding and thus change the core of their response. When students made factor changes, they were more likely to improve their scores from incorrect to correct. This again is parallel to many response change literature for other item types (Jeon et al., 2017; van der Linden & Jeon, 2012).

**Implications**

Furthermore, the insights gained from this study have practical implications for educational practices and assessment designs. For example, by understanding the types of changes that most significantly impact student performance, educators can tailor their feedback and instructional strategies to address these areas specifically. This approach can help improve student learning outcomes by providing more targeted and effective support. For instance, we may find that correcting grammar (like grammatical changes) can have a larger impact on score improvement in extended constructed responses compared to short constructed responses, which could have potential implications for instructional strategies. This work may also help in detecting potential cheating behavior, as unusually high frequencies of certain types of answer changes might indicate aberrant behavior (Jeon et al., 2017; van der Linden & Jeon, 2012). Additionally, insights gained from response change analysis may guide the development of interventions to improve student learning outcomes by addressing common misconceptions or errors identified through their changes in responses.

In terms of the utility of process data, the current study showed the potential to incorporate process data into scoring measures to provide more nuanced interpretations of scores, especially for constructed response items. The use of process data to explore and score intermediate constructed responses provides a path to better understand student scores overall. Using process data in this way also serves as an example of a higher-level use of process data, according to the framework by Bergner & von Davier (2019).

**Limitations and Future Directions**

Although the current study does add to the body of research regarding response change analysis, the use of process data, and machine learning methods, it is not without its limitations. There are some limitations in the analysis and in the framework designed to respond to the two research questions that we want to address and learn from to better navigate future research.

First, although this work focuses on student response change behavior with respect to their writing behaviors, the item we used to analyze this behavior comes from a mathematics assessment. As students' writing skills are not explicitly measured in this selected item or even in the mathematics subject, grammar and coherence in explaining their answer may not fully matter in the final response score. In the results section we noted how grammar and structure changes did not contribute as much as factor change to student scores. If we conducted the same analysis in other assessment subjects where writing skills are more explicitly needed (e.g., reading and writing) we might see variations in the impact of dimension on scores. In future research, we would like to investigate the use of our framework on response change when language and writing have a more significant effect on student scores, such as the NAEP Reading assessment.

Second, analysis in the current study is conducted using process data collected from one item. While the observations made about student response change behavior is consistent with literature from other item types, to make claims about student behavior on constructed response items we must conduct a more comprehensive behavior analysis on other CR items across subjects and years. With our current framework, the ability to analyze other subjects is fairly straightforward; we would only have to train the automated scoring model specifically for each new item. Third, only around 3% of students conducted response changes to the selected item. Learnings from these students may not generalize to the larger population of students. However, this small sample is consistent given that we expect lower response changes to CR items in comparison to other items, as it takes more effort to conduct a dimensional change in CR items.

If one expects to conduct response change analysis with respect to student performance gain/loss they must have the means to obtain scores for students' intermediate responses. An improvement we note for future research is the automated scoring model. Performance metrics depend heavily on how accurate the model is and while the model we used has acceptable performance there may still be better models. Future work will focus on upgrading the model to enhance its performance further. This may include incorporating more sophisticated machine learning techniques, engineering better features and leveraging larger datasets to refine the accuracy and reliability of the classification (Latif & Zhai, 2024;

Morris et al., 2024; Tyack et al., 2024; Whitmer et al., 2023). Specifically, with large language models (LLMs), we could improve performance to accommodate the issues with the partial class classification. Additionally, we would like to integrate the framework into interactive applications, to better visualize the outcomes of dimensional changes. These tools could make it easier to identify key patterns and provide insights into student learning behaviors. We look forward to further investigations to improve in this area.

The framework developed with this work consists of several components which are independent of each other, hence with the development of the field we believe it would also be possible to improve each component in the future. In conclusion, the current study lays the groundwork for a comprehensive framework for analyzing student responses and identifying key patterns in response changes. With continued development and application, our framework holds the promise of significantly advancing our understanding of student learning and student testing behavior to improve educational outcomes across diverse contexts.

**Footnotes**

[1]https://nltk.org/api/nltk.tokenize.word_tokenize.html

**References**


Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing*, *22*(4), 509–531.

Aninditya, A., Hasibuan, M. A., & Sutoyo, E. (2019). Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy. *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 112–117. https://doi.org/10.1109/IoTaIS47347.2019.8980428

Beck, M. D. (1978). The Effect of Item Response Changes on Scores on an Elementary Reading Achievement Test. *The Journal of Educational Research*, *71*(3), 153–156. https://doi.org/10.1080/00220671.1978.10885059

Benjamin, L., Cavell, T., & Shallenberger, W. (1984). Staying with Initial Answers on Objective Tests: Is it a Myth? *Teaching of Psychology*, *11*, 133–141. https://doi.org/10.1177/009862838401100303

Bergner, Y., & von Davier, A. A. (2019). Process Data in NAEP: Past, Present, and Future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732. https://doi.org/10.3102/1076998618784700

Bridgeman, B. (2012). A Simple Answer to a Simple Question on Changing Answers. *Journal of Educational Measurement*, *49*(4), 467–468. https://doi.org/10.1111/j.1745-3984.2012.00189.x

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Engblom, C., Andersson, K., & Åkerlund, D. (2020). Young students making textual changes during digital writing. *Nordic Journal of Digital Literacy*, *15*(3), 190–201. https://doi.org/10.18261/issn.1891-943x-2020-03-05

Ercikan, K., Guo, H., & He, Q. (2020). Use of Response Process Data to Inform Group Comparisons and Fairness Research. *Educational Assessment*, *25*(3), 179–197. https://doi.org/10.1080/10627197.2020.1804353

Hojeij, Z., & Hurley, Z. (2017). The Triple Flip: Using Technology for Peer and Self-Editing of Writing. *International Journal for the Scholarship of Teaching and Learning*, *11*(1). https://eric.ed.gov/?id=EJ1136125

Ivanova, M. G., & Michaelides, M. P. (2023). Measuring Test-Taking Effort on Constructed-Response Items with Item Response Time and Number of Actions. *Practical Assessment, Research, and Evaluation*, *28*(1), Article 1. https://doi.org/10.7275/pare.1921

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling Answer Change Behavior: An Application of a Generalized Item Response Tree Model. *Journal of Educational and Behavioral Statistics*, *42*(4), 467–490. https://doi.org/10.3102/1076998616688015

Johnson, E. G. (1992). The Design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*(2), 95–110. https://doi.org/10.1111/j.1745-3984.1992.tb00369.x

Kim, H.-K., & Kim, H. A. (2022). Analysis of Student Responses to Constructed Response Items in the Science Assessment of Educational Achievement in South Korea. *International Journal of Science & Mathematics Education*, *20*(5), 901–919. https://doi.org/10.1007/s10763-021-10198-7

Kloosterman, P., Mohr, D., & Walcott, C. (2015). *What Mathematics Do Students Know and How is that Knowledge Changing?: Evidence from the National Assessment of Educational Progress*. IAP.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(1), 8. https://doi.org/10.1186/s40536-014-0008-1

Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised general test. *Educational and Psychological Measurement*, *75*(6), 1002–1020.

Malekian, D., Bailey, J., Kennedy, G., de Barba, P., & Nawaz, S. (2019). Characterising Students' Writing Processes Using Temporal Keystroke Analysis. *International Educational Data Mining Society*. https://eric.ed.gov/?id=ED599193

McMorris, R. F., Schwarz, S. P., Richichi, R. V., Fischer, M., Buczek, N. M., Chevalier, C. L., & Meland, K. A. (1991). *Why do young students change answers on tests?* https://eric.ed.gov/?id=ED342803

Moore, S., Nguyen, H. A., & Stamper, J. (2021). Examining the Effects of Student Participation and Performance on the Quality of Learnersourcing Multiple-Choice Questions. *Proceedings of the Eighth ACM Conference on Learning @ Scale*, 209–220. https://doi.org/10.1145/3430895.3460140

Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2024). Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00418-w

*NAEP Process Data*. (n.d.). The Nation's Report Card. Retrieved January 19, 2025, from https://www.nationsreportcard.gov/process_data/

National Center for Education Statistics. (n.d.). *Assessment Frameworks | NAEP*. National Center for Education Statistics. Retrieved January 19, 2025, from https://nces.ed.gov/nationsreportcard/assessments/frameworks.aspx

Ouyang, W., Harik, P., Clauser, B. E., & Paniagua, M. A. (2019). Investigation of answer changes on the USMLE® Step 2 Clinical Knowledge examination. *BMC Medical Education*, *19*(1), 389. https://doi.org/10.1186/s12909-019-1816-3

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

Qiao, X., & Hicks, J. (2020, August 11). *Exploring Answer Change Behavior Using NAEP Process Data*. AIR - Technical Memorandum.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment. *Applied Measurement in Education*, *26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Smith, M. D. (2017). Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes? *American Educational Research Journal*, *54*(6), 1256–1287. https://doi.org/10.3102/0002831217717949

Tate, T. P., & Warschauer, M. (2019). Keypresses and Mouse Clicks: Analysis of the First National Computer-Based Writing Assessment. *Technology, Knowledge and Learning*, *24*(4), 523–543. https://doi.org/10.1007/s10758-019-09412-x

Tiemann, G. (2015). *An Investigation of Answer Changing on a Large-Scale Computer-Based Educational Assessment* [(Doctoral dissertation,]. University of Kansas.

Tyack, L., Khorramdel, L., & von Davier, M. (2024). Using convolutional neural networks to automatically score eight TIMSS 2019 graphical response items. *Computers and Education: Artificial Intelligence*, *6*, 100249. https://doi.org/10.1016/j.caeai.2024.100249

van der Linden, W. J., & Jeon, M. (2012). Modeling Answer Changes on Test Items. *Journal of Educational and Behavioral Statistics*, *37*(1), 180–199. https://doi.org/10.3102/1076998610396899

Whitmer, J., Beiting-Parrish, M., Blankenship, C., Folwer-Dawson, A., & Pitcher, M. (2023). *NAEP Math Item Automated Scoring Data Challenge Results: High Accuracy and Potential for Additional Insights*.